

Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/JP05/000505

International filing date: 17 January 2005 (17.01.2005)

Document type: Certified copy of priority document

Document details: Country/Office: JP
Number: 2004-018715
Filing date: 27 January 2004 (27.01.2004)

Date of receipt at the International Bureau: 10 March 2005 (10.03.2005)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)



World Intellectual Property Organization (WIPO) - Geneva, Switzerland
Organisation Mondiale de la Propriété Intellectuelle (OMPI) - Genève, Suisse

日 本 国 特 許 庁
JAPAN PATENT OFFICE

20.01.2005

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 4 年 1 月 2 7 日
Date of Application:

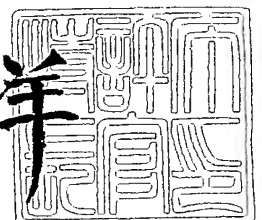
出 願 番 号 特 願 2 0 0 4 - 0 1 8 7 1 5
Application Number:
[ST. 10/C]: [J P 2 0 0 4 - 0 1 8 7 1 5]

出 願 人 松 下 電 器 産 業 株 式 会 社
Applicant(s):

2 0 0 5 年 2 月 2 5 日

特許庁長官
Commissioner,
Japan Patent Office

小 川 洋



【書類名】 特許願
【整理番号】 2033850244
【あて先】 特許庁長官殿
【国際特許分類】 G10L 13/00
【発明者】
 【住所又は居所】 大阪府門真市大字門真 1 0 0 6 番地 松下電器産業株式会社内
 【氏名】 齋藤 夏樹
【発明者】
 【住所又は居所】 大阪府門真市大字門真 1 0 0 6 番地 松下電器産業株式会社内
 【氏名】 釜井 孝浩
【発明者】
 【住所又は居所】 大阪府門真市大字門真 1 0 0 6 番地 松下電器産業株式会社内
 【氏名】 加藤 弓子
【特許出願人】
 【識別番号】 000005821
 【氏名又は名称】 松下電器産業株式会社
【代理人】
 【識別番号】 100109210
 【弁理士】
 【氏名又は名称】 新居 広守
【手数料の表示】
 【予納台帳番号】 049515
 【納付金額】 21,000円
【提出物件の目録】
 【物件名】 特許請求の範囲 1
 【物件名】 明細書 1
 【物件名】 図面 1
 【物件名】 要約書 1
 【包括委任状番号】 0213583

【書類名】 特許請求の範囲**【請求項 1】**

第 1 の声質に属する複数の音声素片に関する第 1 の音声素片情報、及び前記第 1 の声質と異なる第 2 の声質に属する複数の音声素片に関する第 2 の音声素片情報を予め記憶している記憶手段と、

テキストデータを取得するとともに、前記記憶手段の第 1 の音声素片情報から、前記テキストデータに含まれる文字に対応した前記第 1 の声質の合成音声を示す第 1 の合成音声情報を生成し、前記記憶手段の第 2 の音声素片情報から、前記テキストデータに含まれる文字に対応した前記第 2 の声質の合成音声を示す第 2 の合成音声情報を生成する音声情報生成手段と、

前記音声情報生成手段により生成された前記第 1 及び第 2 の合成音声情報から、前記テキストデータに含まれる文字に対応した、前記第 1 及び第 2 の声質の中間的な声質の合成音声を示す中間合成音声情報を生成するモーフィング手段と、

前記モーフィング手段によって生成された前記中間合成音声情報を前記中間的な声質の合成音声に変換して出力する音声出力手段と

を備えることを特徴とする音声合成装置。

【請求項 2】

前記モーフィング手段は、前記音声出力手段から出力される合成音声の声質がその出力中に連続的に変化するように、前記第 1 及び第 2 の合成音声情報の前記中間合成音声情報に対して寄与する割合を変化させる

ことを特徴とする請求項 1 記載の音声合成装置。

【請求項 3】

前記音声情報生成手段は、前記第 1 及び第 2 の合成音声情報を音声生成モデルの特徴パラメタ値列として生成し、

前記モーフィング手段は、前記第 1 及び第 2 の合成音声情報の互いに対応するパラメタの中間値を計算することで、前記中間合成音声情報を生成する

ことを特徴とする請求項 1 又は 2 記載の音声合成装置。

【請求項 4】

前記音声情報生成手段は、前記第 1 及び第 2 の合成音声情報を周波数スペクトルとして生成し、

前記モーフィング手段は、前記第 1 及び第 2 の合成音声情報の周波数スペクトルの中間的な周波数スペクトルを前記中間合成音声情報として生成する

ことを特徴とする請求項 1 又は 2 記載の音声合成装置。

【請求項 5】

前記音声情報生成手段は、前記第 1 及び第 2 の合成音声情報を音声波形として生成し、

前記モーフィング手段は、前記第 1 及び第 2 の合成音声情報の音声波形の中間的な音声波形を前記中間合成音声情報として生成する

ことを特徴とする請求項 1 又は 2 記載の音声合成装置。

【請求項 6】

前記記憶手段は、前記第 1 及び第 2 の音声素片情報のそれぞれにより示される各音声素片における基準を示す内容の特徴情報を、前記第 1 及び第 2 の音声素片情報のそれぞれに含めて記憶しており、

前記音声情報生成手段は、前記第 1 及び第 2 の合成音声情報を、それぞれに前記特徴情報を含めて生成し、

前記モーフィング手段は、前記第 1 及び第 2 の合成音声情報を、それぞれに含まれる前記特徴情報によって示される基準を用いて整合した上で前記中間合成音声情報を生成する

ことを特徴とする請求項 1～5 の何れか 1 項に記載の音声合成装置。

【請求項 7】

前記基準は、前記第 1 及び第 2 の音声素片情報のそれぞれにより示される各音声素片の音響的特徴の変化点である

ことを特徴とする請求項 6 記載の音声合成装置。

【請求項 8】

前記音響的特徴の変化点は、前記第 1 及び第 2 の音声素片情報のそれぞれに示される各音声素片を HMM (Hidden Markov Model) で表した最尤経路上の状態遷移点であって、

前記モーフィング手段は、前記第 1 及び第 2 の合成音声情報を、前記状態遷移点を用いて時間軸上で整合した上で前記中間合成音声情報を生成する

ことを特徴とする請求項 7 記載の音声合成装置。

【請求項 9】

前記基準は、前記第 1 及び第 2 の音声素片情報のそれぞれにより示される各音声素片の、周波数スペクトルに現れる特徴的な周波数及び時間の関数であって、

前記モーフィング手段は、前記第 1 及び第 2 の合成音声情報を、前記関数を用いて周波数軸上で整合した上で前記中間合成音声情報を生成する

ことを特徴とする請求項 6 記載の音声合成装置。

【請求項 10】

前記音声合成装置は、さらに、

前記第 1 の声質に対応する画像を示す第 1 の画像情報、及び前記第 2 の声質に対応する画像を示す第 2 の画像情報を予め記憶している画像記憶手段と、

前記第 1 及び第 2 の画像情報のそれぞれにより示される画像の中間的な画像であって、前記中間合成音声情報の声質に対応する画像を示す中間画像情報を、前記第 1 及び第 2 の画像情報から生成する画像モーフィング手段と、

前記画像モーフィング手段により生成された中間画像情報を取得して、前記中間画像情報により示される画像を、前記音声出力手段から出力される合成音声に同期させて表示する表示手段と

を備えることを特徴とする請求項 1～9 の何れか 1 項に記載の音声合成装置。

【請求項 11】

前記第 1 の画像情報は前記第 1 の声質に対応する顔画像を示し、前記第 2 の画像情報は前記第 2 の声質に対応する顔画像を示す

ことを特徴とする請求項 10 記載の音声合成装置。

【請求項 12】

前記音声合成装置は、さらに、

前記第 1 及び第 2 の声質を示す固定点、及びユーザの操作に基づいて移動する移動点をそれぞれ N 次元 (N は自然数) の座標上に配置して表し、前記固定点及び移動点の配置に基づいて、前記第 1 及び第 2 の合成音声情報の前記中間合成音声情報に対して寄与する割合を導出し、導出した割合を前記モーフィング手段に指示する指定手段を備え、

前記モーフィング手段は、前記指定手段により指定された割合に応じて、前記中間合成音声情報を生成する

ことを特徴とする請求項 1～11 の何れか 1 項に記載の音声合成装置。

【請求項 13】

前記音声情報生成手段は、

前記第 1 及び第 2 の合成音声情報のそれぞれを順次生成する

ことを特徴とする請求項 1～12 の何れか 1 項に記載の音声合成装置。

【請求項 14】

前記音声情報生成手段は、

前記第 1 及び第 2 の合成音声情報のそれぞれを並列に生成する

ことを特徴とする請求項 1～12 の何れか 1 項に記載の音声合成装置。

【請求項 15】

第 1 の声質に属する複数の音声素片に関する第 1 の音声素片情報、及び前記第 1 の声質と異なる第 2 の声質に属する複数の音声素片に関する第 2 の音声素片情報を予め記憶しているメモリを用いることで、合成音声を生成して出力する音声合成方法であって、

テキストデータを取得するテキスト取得ステップと、

前記メモリの第1の音声素片情報から、前記テキストデータに含まれる文字に対応した前記第1の声質の合成音声を示す第1の合成音声情報を生成し、前記メモリの第2の音声素片情報から、前記テキストデータに含まれる文字に対応した前記第2の声質の合成音声を示す第2の合成音声情報を生成する音声情報生成ステップと、

前記音声情報生成ステップで生成された前記第1及び第2の合成音声情報から、前記テキストデータに含まれる文字に対応した、前記第1及び第2の声質の中間的な声質の合成音声を示す中間合成音声情報を生成するモーフィングステップと、

前記モーフィングステップで生成された前記中間合成音声情報を前記中間的な声質の合成音声に変換して出力する音声出力ステップと

を含むことを特徴とする音声合成方法。

【請求項16】

前記モーフィングステップでは、前記音声出力ステップで出力される合成音声の声質がその出力中に連続的に変化するように、前記第1及び第2の合成音声情報の前記中間合成音声情報に対して寄与する割合を変化させる

ことを特徴とする請求項15記載の音声合成方法。

【請求項17】

前記音声情報生成ステップでは、前記第1及び第2の合成音声情報を音声生成モデルの特徴パラメタ値列として生成し、

前記モーフィングステップでは、前記第1及び第2の合成音声情報の互いに対応するパラメタの中間値を計算することで、前記中間合成音声情報を生成する

ことを特徴とする請求項15又は16記載の音声合成方法。

【請求項18】

前記音声合成方法は、さらに、

前記第1の声質に対応する画像を示す第1の画像情報、及び前記第2の声質に対応する画像を示す第2の画像情報を予め記憶している画像メモリを用い、

前記第1及び第2の画像情報のそれぞれにより示される画像の中間的な画像であって、前記中間合成音声情報の声質に対応する画像を示す中間画像情報を、前記画像メモリの第1及び第2の画像情報から生成する画像モーフィングステップと、

前記画像モーフィングステップで生成された中間画像情報により示される画像を、前記音声出力ステップで出力される合成音声に同期させて表示する表示ステップと

を含むことを特徴とする請求項15～17の何れか1項に記載の音声合成方法。

【請求項19】

前記第1の画像情報は前記第1の声質に対応する顔画像を示し、前記第2の画像情報は前記第2の声質に対応する顔画像を示す

ことを特徴とする請求項18記載の音声合成方法。

【請求項20】

請求項15～19の何れか1項に記載の音声合成方法に含まれるステップをコンピュータに実行させるプログラム。

【書類名】明細書

【発明の名称】音声合成装置

【技術分野】

【0001】

本発明は、合成音声を生成して出力する音声合成装置に関する。

【背景技術】

【0002】

従来より、所望の合成音声を生成して出力する音声合成装置が提供されている（例えば、特許文献1及び特許文献2参照。）。

特許文献1の音声合成装置は、それぞれ声質の異なる複数の音声素片データベースを備え、これらの音声素片データベースを切り替えて用いることにより、所望の合成音声を生成して出力する。

【0003】

また、特許文献2の音声合成装置（音声変形装置）は、音声分析結果のスペクトルパラメタを変換することにより、所望の合成音声を生成して出力する。

【特許文献1】特開平7-319495号公報

【特許文献2】特開2000-330582号公報

【発明の開示】

【発明が解決しようとする課題】

【0004】

しかしながら、上記特許文献1及び特許文献2の音声合成装置では、声質変換の自由度が狭かったり、音質の調整が非常に困難であるという問題がある。

即ち、特許文献1では、合成音声の声質が予め設定された声質に限られ、その予め設定された声質間の連続的な変化を表現することができない。

また、特許文献2では、特に音声の基本周波数や音源パワーなどの特徴パラメタのダイナミックレンジを大きくしてしまうと音質に破綻が生じてしまい、良い音質を維持するのが困難となる。

【0005】

そこで、本発明は、このような問題に鑑みてなされたものであって、声質の自由度が広く良い音質の合成音声テキストデータから生成する音声合成装置を提供することを目的とする。

【課題を解決するための手段】

【0006】

上記目的を達成するために、本発明に係る音声合成装置は、第1の声質に属する複数の音声素片に関する第1の音声素片情報、及び前記第1の声質と異なる第2の声質に属する複数の音声素片に関する第2の音声素片情報を予め記憶している記憶手段と、テキストデータを取得するとともに、前記記憶手段の第1の音声素片情報から、前記テキストデータに含まれる文字に対応した前記第1の声質の合成音声を示す第1の合成音声情報を生成し、前記記憶手段の第2の音声素片情報から、前記テキストデータに含まれる文字に対応した前記第2の声質の合成音声を示す第2の合成音声情報を生成する音声情報生成手段と、前記音声情報生成手段により生成された前記第1及び第2の合成音声情報から、前記テキストデータに含まれる文字に対応した、前記第1及び第2の声質の中間的な声質の合成音声を示す中間合成音声情報を生成するモーフィング手段と、前記モーフィング手段によって生成された前記中間合成音声情報を前記中間的な声質の合成音声に変換して出力する音声出力手段とを備えることを特徴とする。例えば、前記音声情報生成手段は、前記第1及び第2の合成音声情報を音声生成モデルの特徴パラメタ値列として生成し、前記モーフィング手段は、前記第1及び第2の合成音声情報の互に対応するパラメタの中間値を計算することで、前記中間合成音声情報を生成する。

【0007】

これにより、第1の声質に対する第1の音声素片情報、及び第2の声質に対する第2の

音声素片情報だけを記憶手段に予め記憶させておけば、第1及び第2の声質の中間的な声質の合成音声出力されるため、記憶手段に予め記憶しておく内容の声質に限定されず、声質の自由度を広めることができる。また、第1及び第2の声質を有する第1及び第2の合成音声情報を基礎に中間合成音声情報が生成されるため、従来例のように基本周波数などのパラメタのダイナミックレンジを大きくしすぎるような処理がなされず、合成音声の音質を良い状態に維持することができる。また、本発明に係る音声合成装置は、テキストデータを取得して、そこに含まれる文字列に応じた合成音声出力するため、ユーザに対する使い勝手を向上することができる。

【0008】

ここで、前記モーフィング手段は、前記音声出力手段から出力される合成音声の声質がその出力中に連続的に変化するように、前記第1及び第2の合成音声情報の前記中間合成音声情報に対して寄与する割合を変化させることを特徴としても良い。

これにより、合成音声の出力中にその合成音声の声質が連続的に変化するため、例えば、平常声から怒り声に連続的に変化するような合成音声出力することができる。

【0009】

また、前記記憶手段は、前記第1及び第2の音声素片情報のそれぞれにより示される各音声素片における基準を示す内容の特徴情報を、前記第1及び第2の音声素片情報のそれぞれに含めて記憶しており、前記音声情報生成手段は、前記第1及び第2の合成音声情報を、それぞれに前記特徴情報を含めて生成し、前記モーフィング手段は、前記第1及び第2の合成音声情報を、それぞれに含まれる前記特徴情報によって示される基準を用いて整合した上で前記中間合成音声情報を生成することを特徴としても良い。例えば、前記基準は、前記第1及び第2の音声素片情報のそれぞれにより示される各音声素片の音響的特徴の変化点である。

【0010】

これにより、モーフィング手段による中間合成音声情報の生成に、第1及び第2の合成音声情報が上述の基準を用いて整合されるため、例えば第1及び第2の合成音声情報をパターンマッチングなどによって整合するような場合と比べ、迅速に整合を図って中間合成音声情報を生成することができ、その結果、処理速度を向上することができる。

【0011】

また、前記音声合成装置は、さらに、前記第1の声質に対応する画像を示す第1の画像情報、及び前記第2の声質に対応する画像を示す第2の画像情報を予め記憶している画像記憶手段と、前記第1及び第2の画像情報のそれぞれにより示される画像の中間的な画像であって、前記中間合成音声情報の声質に対応する画像を示す中間画像情報を、前記第1及び第2の画像情報から生成する画像モーフィング手段と、前記画像モーフィング手段により生成された中間画像情報を取得して、前記中間画像情報により示される画像を、前記音声出力手段から出力される合成音声に同期させて表示する表示手段とを備えることを特徴としても良い。例えば、前記第1の画像情報は前記第1の声質に対応する顔画像を示し、前記第2の画像情報は前記第2の声質に対応する顔画像を示す。

【0012】

これにより、第1及び第2の声質の中間的な声質に対応する顔画像が、その中間的な声質の合成音声の出力と同期して表示されるため、合成音声の声質を顔画像の表情からもユーザに伝えることができ、表現力の向上を図ることができる。

【0013】

ここで、前記音声情報生成手段は、前記第1及び第2の合成音声情報のそれぞれを順次生成することを特徴としても良い。

これにより、音声情報生成手段の単位時間あたりの処理負担を軽減することができ、音声情報生成手段の構成を簡単にするができる。その結果、装置全体を小型化することができるとともに、コスト低減を図ることができる。

【0014】

また、前記音声情報生成手段は、前記第1及び第2の合成音声情報のそれぞれを並列に

生成することを特徴としても良い。

これにより、第1及び第2の合成音声情報を迅速に生成することができ、その結果、テキストデータの取得から合成音声の出力までの時間を短縮することができる。

なお、本発明は、上述の音声合成装置の合成音声を生成して出力する方法やプログラム、そのプログラムを格納する記憶媒体としても実現することができる。

【発明の効果】

【0015】

本発明の音声合成装置では、声質の自由度が広く良い音質の合成音声テキストデータから生成することができるという効果を奏する。

【発明を実施するための最良の形態】

【0016】

以下、本発明の実施の形態について図面を用いて詳細に説明する。

（実施の形態1）

図1は、本発明の実施の形態1に係る音声合成装置の構成を示す構成図である。

本実施の形態の音声合成装置は、声質の自由度が広く良い音質の合成音声テキストデータから生成するものであって、複数の音声素片（音素）に関する音声素片データを蓄積する複数の音声合成DB101a～101zと、1つの音声合成DBに蓄積された音声素片データを用いることにより、テキスト10に示される文字列に対応する音声合成パラメタ値列11を生成する複数の音声合成部（音声情報生成手段）103と、ユーザによる操作に基づいて声質を指定する声質指定部104と、複数の音声合成部103により生成された音声合成パラメタ値列11を用いて音声モーフィング処理を行い、中間的合成音波形データ12を出力する音声モーフィング部105と、中間的合成音波形データ12に基づいて合成音声出力するスピーカ107とを備えている。

【0017】

音声合成DB101a～101zのそれぞれが蓄積する音声素片データの示す声質は異なっている。例えば、音声合成DB101aには、笑っている声質の音声素片データが蓄積され、音声合成DB101zには、怒っている声質の音声素片データが蓄積されている。また、本実施の形態における音声素片データは、音声生成モデルの特徴パラメタ値列の形式で表現されている。さらに、蓄積される各音声素片データには、これらのデータにより示される各音声素片の開始及び終了の時刻と、音響的特徴の変化点の時刻とを示すラベル情報が付されている。

【0018】

複数の音声合成部103は、それぞれ上述の音声合成DBと一対一に対応付けられている。このような音声合成部103の動作について図2を参照して説明する。

図2は、音声合成部103の動作を説明するための説明図である。

音声合成部103は、図2に示すように、言語処理部103aと素片結合部103bとを備えている。

【0019】

言語処理部103aは、テキスト10を取得して、テキスト10に示される文字列を音素情報10aに変換する。音素情報10aは、テキスト10に示される文字列が音素列の形で表現されたもので、他にアクセント位置情報や音素継続長情報など、素片選択・結合・変形に必要な情報を含んでもよい。

素片結合部103bは、対応付けられた音声合成DBの音声素片データから適切な音声素片に関する部分を抜き出して、抜き出した部分の結合と変形を行うことにより、言語処理部103aにより出力される音素情報10aに対応する音声合成パラメタ値列11を生成する。音声合成パラメタ値列11は、実際の音声波形を生成するために必要となる十分な情報を含んだ複数の特徴パラメタの値が配列されたものである。例えば、音声合成パラメタ値列11は、時系列に沿った各音声分析合成フレームごとに、図2に示すような、5つの特徴パラメタを含んで構成される。5つの特徴パラメタとは、音声の基本周波数F0と、第一フォルマントF1と、第二フォルマントF2と、音声分析合成フレーム継続長F

Rと、音源強度PWとである。また、上述のように音声素片データにはラベル情報が付されているので、このように生成される音声合成パラメタ値列11にもラベル情報が付されている。

【0020】

声質指定部104は、ユーザによる操作に基づき、何れの音声合成パラメタ値列11を用い、その音声合成パラメタ値列11に対してどのような割合で音声モーフィング処理を行うかを音声モーフィング部105に指示する。さらに、声質指定部104はその割合を時系列に沿って変化させる。このような声質指定部104は、例えばパーソナルコンピュータなどから構成され、ユーザにより操作された結果を表示するディスプレイを備えている。

【0021】

図3は、声質指定部104のディスプレイが表示する画面の一例を示す画面表示図である。

ディスプレイには、音声合成部DB101a~101zの声質を示す複数の声質アイコンが表示されている。なお図3では、複数の声質アイコンのうち、声質Aの声質アイコン104Aと、声質Bの声質アイコン104Bと、声質Zの声質アイコン104Zとを示す。このような複数の声質アイコンは、それぞれの示す声質が似ているものほど互いに近寄るように配置され、似ていないものほど互いに離れるように配置される。

【0022】

ここで、声質指定部104は、このようなディスプレイ上に、ユーザによる操作に応じて移動可能な指定アイコン104iを表示する。

声質指定部104は、ユーザによって配置された指定アイコン104iから近い声質アイコンを調べ、例えば声質アイコン104A、104B、104Zを特定すると、声質Aの音声合成パラメタ値列11と、声質Bの音声合成パラメタ値列11と、声質Zの音声合成パラメタ値列11とを用いることを、音声モーフィング部105に指示する。さらに、声質指定部104は、各声質アイコン104A、104B、104Z及び指定アイコン104iの相対的な配置に対応する割合を、音声モーフィング部105に指示する。

【0023】

即ち、声質指定部104は、指定アイコン104iから各声質アイコン104A、104B、104Zまでの距離を調べ、それらの距離に応じた割合を指示する。

又は、声質指定部104は、まず、声質Aと声質Zの中間的な声質（テンポラリ声質）を生成するための割合を求め、次に、そのテンポラリ声質と声質Bとから、指定アイコン104iで示される声質を生成するための割合を求め、これらの割合を指示する。具体的に、声質指定部104は、声質アイコン104A及び声質アイコン104Zを結ぶ直線と、声質アイコン104B及び指定アイコン104iを結ぶ直線とを算出し、これらの直線の交点の位置104tを特定する。この位置104tにより示される声質が上述のテンポラリ声質である。そして、声質指定部104は、位置104tから各声質アイコン104A、104Zまでの距離の割合を求める。次に、声質指定部104は、指定アイコン104iから声質アイコン104B及び位置104tまでの距離の割合を求め、このように求めた2つの割合を指示する。

【0024】

このような声質指定部104を操作することにより、ユーザは、スピーカ107から出力させようとする合成音声の声質の、予め設定された声質に対する類似度を容易に入力することができる。そこでユーザは、例えば声質Aに近い合成音声をスピーカ107から出力させたいときには、指定アイコン104iが声質アイコン104Aに近づくように声質指定部104を操作する。

また、声質指定部104は、ユーザからの操作に応じて、上述のような割合を時系列に沿って連続的に変化させる。

【0025】

図4は、声質指定部104のディスプレイが表示する他の画面の一例を示す画面表示図

である。

声質指定部104は、図4に示すように、ユーザによる操作に応じて、ディスプレイ上に3つのアイコン21, 22, 23を配置し、アイコン21からアイコン22を通してアイコン23に到達するような軌跡を特定する。そして、声質指定部104は、その軌跡に沿って指定アイコン104iが移動するように、上述の割合を時系列に沿って連続的に変化させる。例えば、声質指定部104は、その軌跡の長さをLとすると、毎秒 $0.01 \times L$ の速度で指定アイコン104iが移動するように、その割合を変化させる。

音声モーフィング部105は、上述のような声質指定部104により指定された音声合成パラメタ値列11と割合とから、音声モーフィング処理を行う。

【0026】

図5は、音声モーフィング部105の処理動作を説明するための説明図である。

音声モーフィング部105は、図5に示すように、パラメタ中間値計算部105aと、波形生成部105bとを備えている。

【0027】

パラメタ中間値計算部105aは、声質指定部104により指定された少なくとも2つの音声合成パラメタ値列11と割合とを特定し、それらの音声合成パラメタ値列11から、互いに対応する音声分析合成フレーム間ごとに、その割合に応じた中間的音声合成パラメタ値列13を生成する。

例えば、パラメタ中間値計算部105aは、声質指定部104の指定に基づいて、声質Aの音声合成パラメタ値列11と、声質Zの音声合成パラメタ値列11と、割合50:50とを特定すると、まず、その声質Aの音声合成パラメタ値列11と、声質Zの音声合成パラメタ値列11とを、それぞれに対応する音声合成部103から取得する。そして、パラメタ中間値計算部105aは、互いに対応する音声分析合成フレームにおいて、声質Aの音声合成パラメタ値列11に含まれる各特徴パラメタと、声質Zの音声合成パラメタ値列11に含まれる各特徴パラメタとの中間値を50:50の割合で算出し、その算出結果を中間的音声合成パラメタ値列13として生成する。具体的に、互いに対応する音声分析合成フレームにおいて、声質Aの音声合成パラメタ値列11の基本周波数F0の値が300であり、声質Zの音声合成パラメタ値列11の基本周波数F0の値が280である場合には、パラメタ中間値計算部105aは、当該音声分析合成フレームでの基本周波数F0が290となる中間的音声合成パラメタ値列13を生成する。

【0028】

また、図3を用いて説明したように、声質指定部104により、声質Aの音声合成パラメタ値列11と、声質Bの音声合成パラメタ値列11と、声質Zの音声合成パラメタ値列11とが指定され、さらに、声質Aと声質Zの中間的なテンポラリ声質を生成するための割合（例えば3:7）と、そのテンポラリ声質と声質Bとから指定アイコン104iで示される声質を生成するための割合（例えば9:1）とが指定され場合には、音声モーフィング部105は、まず、声質Aの音声合成パラメタ値列11と、声質Zの音声合成パラメタ値列11とを用いて、3:7の割合に応じた音声モーフィング処理を行う。これにより、テンポラリ声質に対応する音声合成パラメタ値列が生成される。さらに、音声モーフィング部105は、先に生成した音声合成パラメタ値列と、声質Bの音声合成パラメタ値列11とを用いて、9:1の割合に応じた音声モーフィング処理を行う。これにより、指定アイコン104iに対応する中間的音声合成パラメタ値列13が生成される。ここで、上述の3:7の割合に応じた音声モーフィング処理とは、声質Aの音声合成パラメタ値列11を $3 / (3 + 7)$ だけ声質Zの音声合成パラメタ値列11に近づける処理であり、逆に、声質Zの音声合成パラメタ値列11を $7 / (3 + 7)$ だけ声質Aの音声合成パラメタ値列11に近づける処理をいう。この結果、生成される音声合成パラメタ値列は、声質Zの音声合成パラメタ値列11よりも、声質Aの音声合成パラメタ値列11に類似することとなる。

【0029】

波形生成部105bは、パラメタ中間値計算部105aにより生成された中間的音声合

成パラメタ値列 13 を取得して、その中間的音声合成パラメタ値列 13 に応じた中間的合成音波形データ 12 を生成し、スピーカ 107 に対して出力する。

これにより、スピーカ 107 からは、中間的音声合成パラメタ値列 13 に応じた合成音声出力される。即ち、予め設定された複数の声質の中間的な声質の合成音声はスピーカ 107 から出力される。

【0030】

ここで、一般に複数の音声合成パラメタ値列 11 に含まれる音声分析合成フレームの総数はそれぞれ異なるため、パラメタ中間値計算部 105a は、上述のように互いに異なる声質の音声合成パラメタ値列 11 を用いて音声モーフィング処理を行うときには、音声分析合成フレーム間の対応付けを行うために時間軸アライメントを行う。

即ちパラメタ中間値計算部 105a は、音声合成パラメタ値列 11 に付されたラベル情報に基づいて、これらの音声合成パラメタ値列 11 の時間軸上の整合を図る。

【0031】

ラベル情報は、前述のように各音声素片の開始及び終了の時刻と、音響的特徴の変化点の時刻とを示す。音響的特徴の変化点は、例えば、音声素片に対応する不特定話者 HMM 音素モデルにより示される最尤パスの状態遷移点である。

図 6 は、音声素片と HMM 音素モデルの一例を示す例示図である。

例えば、図 6 に示すように、所定の音声素片 30 を不特定話者 HMM 音素モデル（以下、音素モデルと略す）31 で認識した場合、その音素モデル 31 は、開始状態（ S_0 ）と終了状態（ S_E ）を含めて 4 つの状態（ S_0 , S_1 , S_2 , S_E ）で構成される。ここで、最尤パスの形状 32 は、時刻 4 から 5 において、状態 S_1 から状態 S_2 への状態遷移を有する。つまり、音声合成 DB 101a ~ 101z に格納されている音声素片データの音声素片 30 に対応する部分には、この音声素片 30 の開始時刻 1、終了時刻 N、及び音響的特徴の変化点の時刻 5 を示すラベル情報が付されている。

【0032】

したがって、パラメタ中間値計算部 105a は、そのラベル情報に示される開始時刻 1、終了時刻 N、及び音響的特徴の変換点の時刻 5 に基づいて、時間軸の伸縮処理を行う。即ち、パラメタ中間値計算部 105a は、取得した各音声合成パラメタ値列 11 に対して、ラベル情報により示される時刻が一致するように、その時刻間を線形に伸縮する。

これにより、パラメタ中間値計算部 105a は、各音声合成パラメタ値列 11 に対して、それぞれの音声分析合成フレームの対応付けを行うことができる。つまり、時間軸アライメントを行うことができる。また、このように本実施の形態ではラベル情報を用いて時間軸アライメントを行うことにより、例えば各音声合成パラメタ値列 11 のパターンマッチングなどにより時間軸アライメントを行う場合と比べて、迅速に時間軸アライメントを実行することができる。

【0033】

以上のように本実施の形態では、パラメタ中間値計算部 105a が、声質指定部 104 から指示された複数の音声合成パラメタ値列 11 に対して、声質指定部 104 から指定された割合に応じた音声モーフィング処理を実行するため、合成音声の声質の自由度を広めることができる。

例えば、図 3 に示す声質指定部 104 のディスプレイ上で、ユーザが声質指定部 104 を操作することにより指定アイコン 104i を声質アイコン 104A、声質アイコン 104B 及び声質アイコン 104Z に近づければ、音声モーフィング部 105 は、声質 A の音声合成 DB 101a に基づいて音声合成部 103 により生成された音声合成パラメタ値列 11 と、声質 B の音声合成 DB 101b に基づいて音声合成部 103 により生成された音声合成パラメタ値列 11 と、声質 Z の音声合成 DB 101z に基づいて音声合成部 103 により生成された音声合成パラメタ値列 11 とを用いて、それぞれを同じ割合で音声モーフィング処理する。その結果、スピーカ 107 から出力される合成音声を、声質 A と声質 B と声質 C との中間的な声質にすることができる。また、ユーザが声質指定部 104 を操作することにより指定アイコン 104i を声質アイコン 104A に近づければ、スピーカ

107から出力される合成音声の声質を声質Aに近づけることができる。

【0034】

また、本実施の形態の声質指定部104は、ユーザによる操作に応じてその割合を時系列に沿って変化させるため、スピーカ107から出力される合成音声の声質を時系列に沿ってなめらかに変化させることができる。例えば、図4で説明したように、声質指定部104が、毎秒 $0.01 \times L$ の速度で軌跡上を指定アイコン104iが移動するように割合を変化させた場合には、100秒間声質がなめらかに変化し続けるような合成音声スピーカ107から出力される。

【0035】

これによって、例えば「喋り始めは冷静だが、喋りながら段々怒っていく」というような、従来は不可能だった、表現力の高い音声合成装置が実現できる。また、合成音声の声質を1発声の中で連続的に変化させることもできる。

さらに、本実施の形態では、音声モーフィング処理を行うため、従来例のように声質に破綻が起ることがなく合成音声の品質を維持することができる。

【0036】

なお、本実施の形態では、複数の音声合成部103のそれぞれに音素情報10a及び音声合成パラメタ値列11を生成させたが、音声モーフィング処理に必要な声質に対応する音素情報10aが何れも同じであるときには、1つの音声合成部103の言語処理部103aにのみ音素情報10aを生成させ、その音素情報10aから音声合成パラメタ値列11を生成する処理を、複数の音声合成部103の素片結合部103bにさせても良い。

【0037】

(変形例)

ここで、本実施の形態における音声合成部に関する変形例について説明する。

図7は、本変形例に係る音声合成装置の構成を示す構成図である。

本変形例に係る音声合成装置は、互いに異なる声質の音声合成パラメタ値列11を生成する1つの音声合成部103cを備える。

【0038】

この音声合成部103cは、テキスト10を取得して、テキスト10に示される文字列を音素情報10aに変換した後、複数の音声合成DB101a～101zを順番に切り替えて参照ことで、その音素情報10aに対応する複数の声質の音声合成パラメタ値列11を順次生成する。

音声モーフィング部105は、必要な音声合成パラメタ値列11が生成されるまで待機し、その後、上述と同様の方法で中間的合成音波形データ12を生成する。

【0039】

なお、上述のような場合、声質指定部104は、音声合成部103cに指示して、音声モーフィング部105が必要とする音声合成パラメタ値列11のみを生成させることで、音声モーフィング部105の待機時間を短くすることができる。

このように本変形例では、音声合成部103cを1つだけ備えることにより、音声合成装置全体の小型化並びにコスト低減を図ることができる。

【0040】

(実施の形態2)

図8は、本発明の実施の形態2に係る音声合成装置の構成を示す構成図である。

本実施の形態の音声合成装置は、実施の形態1の音声合成パラメタ値列11の代わりに周波数スペクトルを用い、この周波数スペクトルによる音声モーフィング処理を行う。

このような音声合成装置は、複数の音声素片に関する音声素片データを蓄積する複数の音声合成DB201a～201zと、1つの音声合成DBに蓄積された音声素片データを用いることにより、テキスト10に示される文字列に対応する合成音スペクトル41を生成する複数の音声合成部203と、ユーザによる操作に基づいて声質を指定する声質指定部104と、複数の音声合成部203により生成された合成音スペクトル41を用いて音

声モーフィング処理を行い、中間的合成音波形データ12を出力する音声モーフィング部205と、中間的合成音波形データ12に基づいて合成音声出力するスピーカ107とを備えている。

【0041】

複数の音声合成DB201a~201zのそれぞれが蓄積する音声素片データの示す声質は、実施の形態1の音声合成DB101a~101zと同様、異っている。また、本実施の形態における音声素片データは、周波数スペクトルの形式で表現されている。

複数の音声合成部203は、それぞれ上述の音声合成DBと一対一に対応付けられている。そして、各音声合成部203は、テキスト10を取得して、テキスト10に示される文字列を音素情報に変換する。さらに、音声合成部203は、対応付けられた音声合成DBの音声素片データから適切な音声素片に関する部分を抜き出して、抜き出した部分の結合と変形を行うことにより、先に生成した音素情報に対応する周波数スペクトルたる合成音スペクトル41を生成する。このような合成音スペクトル41は、音声のフーリエ解析結果の形式であっても良く、音声のケプストラムパラメタ値を時系列的に並べた形式であっても良い。

【0042】

声質指定部104は、実施の形態1と同様、ユーザによる操作に基づき、何れの合成音スペクトル41を用い、その合成音スペクトル41に対してどのような割合で音声モーフィング処理を行うかを音声モーフィング部205に指示する。さらに、声質指定部104はその割合を時系列に沿って変化させる。

本実施の形態における音声モーフィング部205は、複数の音声合成部203から出力される合成音スペクトル41を取得して、その中間的性質を持つ合成音スペクトルを生成し、さらに、その中間的性質の合成音スペクトルを中間的合成音波形データ12に変形して出力する。

【0043】

図9は、本実施の形態における音声モーフィング部205の処理動作を説明するための説明図である。

音声モーフィング部205は、図9に示すように、スペクトルモーフィング部205aと、波形生成部205bとを備えている。

スペクトルモーフィング部205aは、声質指定部104により指定された少なくとも2つの合成音スペクトル41と割合とを特定し、それらの合成音スペクトル41から、その割合に応じた中間的合成音スペクトル42を生成する。

【0044】

即ち、スペクトルモーフィング部205aは、複数の合成音スペクトル41から、声質指定部104により指定された2つ以上の合成音スペクトル41を選択する。そして、スペクトルモーフィング部205aは、それら合成音スペクトル41の形状の特徴を示すフォルマント形状50を抽出して、そのフォルマント形状50ができるだけ一致するような変形を各合成音スペクトル41に加えた後、各合成音スペクトル41の重ね合わせを行う。なお、上述の合成音スペクトル41の形状の特徴は、フォルマント形状でなくても良く、例えばある程度以上強く現れていて、かつその軌跡が連続的に追えるものであれば良い。図9に示されるように、フォルマント形状50は、声質Aの合成音スペクトル41及び声質Zの合成音スペクトル41のそれぞれについてスペクトル形状の特徴を模式的に表すものである。

【0045】

具体的に、スペクトルモーフィング部205aは、声質指定部104からの指定に基づき、声質A及び声質Zの合成音スペクトル41と4:6の割合とを特定すると、まず、その声質Aの合成音スペクトル41と声質Zの合成音スペクトル41とを取得して、それらの合成音スペクトル41からフォルマント形状50を抽出する。次に、スペクトルモーフィング部205aは、声質Aの合成音スペクトル41のフォルマント形状50が声質Zの合成音スペクトル41のフォルマント形状50に40%だけ近づくように、声質Aの合成

音スペクトル 41 を周波数軸及び時間軸上で伸縮処理する。さらに、スペクトルモーフィング部 205a は、声質 Z の合成音スペクトル 41 のフォルマント形状 50 が声質 A の合成音スペクトル 41 のフォルマント形状 50 に 60% だけ近づくように、声質 Z の合成音スペクトル 41 を周波数軸及び時間軸上で伸縮処理する。最後に、スペクトルモーフィング部 205a は、伸縮処理された声質 A の合成音スペクトル 41 のパワーを 60% にするとともに、伸縮処理された声質 Z の合成音スペクトル 41 のパワーを 40% にした上で、両合成音スペクトル 41 を重ね合わせる。その結果、声質 A の合成音スペクトル 41 と声質 Z の合成音スペクトル 41 との音声モーフィング処理が 4:6 の割合で行われ、中間的合成音スペクトル 42 が生成される。

【0046】

このような、中間的合成音スペクトル 42 を生成する音声モーフィング処理について、図 10～図 12 を用いてより詳細に説明する。

図 10 は、声質 A 及び声質 Z の合成音スペクトル 41 と、それらに対応する短時間フーリエスペクトルとを示す図である。

スペクトルモーフィング部 205a は、声質 A の合成音スペクトル 41 と声質 Z の合成音スペクトル 41 との音声モーフィング処理を 4:6 の割合で行うときには、まず、上述のようにこれらの合成音スペクトル 41 のフォルマント形状 50 を互いに近づけるため、各合成音スペクトル 41 同士の時間軸アライメントを行う。このような時間軸アライメントは、各合成音スペクトル 41 のフォルマント形状 50 同士のパターンマッチングを行うことにより実現される。なお、各合成音スペクトル 41 もしくはフォルマント形状 50 に関する他の特徴量を用いてパターンマッチングを行ってもよい。

【0047】

即ち、スペクトルモーフィング部 205a は、図 10 に示すように、両合成音スペクトル 41 のそれぞれのフォルマント形状 50 において、パターンが一致するフーリエスペクトル分析窓 51 の部位で時刻が一致するように、両合成音スペクトル 41 に対して時間軸上の伸縮を行う。これにより時間軸アライメントが実現される。

また、図 10 に示すように、互いにパターンが一致するフーリエスペクトル分析窓 51 のそれぞれの短時間フーリエスペクトル 41a には、フォルマント形状 50 の周波数 50a, 50b が互いに異なるように表示される。

【0048】

そこで、時間軸アライメントの完了後、スペクトルモーフィング部 205a は、アライメントされた音声の各時刻において、フォルマント形状 50 を基に、周波数軸上の伸縮処理を行う。即ち、スペクトルモーフィング部 205a は、各時刻における声質 A 及び声質 B の短時間フーリエスペクトル 41a において周波数 50a, 50b が一致するように、両短時間フーリエスペクトル 41a を周波数軸上で伸縮する。

【0049】

図 11 は、スペクトルモーフィング部 205a が両短時間フーリエスペクトル 41a を周波数軸上で伸縮する様子を説明するための説明図である。

スペクトルモーフィング部 205a は、声質 A の短時間フーリエスペクトル 41a 上の周波数 50a, 50b が 40% だけ、声質 Z の短時間フーリエスペクトル 41a 上の周波数 50a, 50b に近づくように、声質 A の短時間フーリエスペクトル 41a を周波数軸上で伸縮し、中間的な短時間フーリエスペクトル 41b を生成する。これと同様に、スペクトルモーフィング部 205a は、声質 Z の短時間フーリエスペクトル 41a 上の周波数 50a, 50b が 60% だけ、声質 A の短時間フーリエスペクトル 41a 上の周波数 50a, 50b に近づくように、声質 Z の短時間フーリエスペクトル 41a を周波数軸上で伸縮し、中間的な短時間フーリエスペクトル 41b を生成する。その結果、中間的な両短時間フーリエスペクトル 41b において、フォルマント形状 50 の周波数は周波数 f1, f2 に揃えられた状態となる。

【0050】

例えば、声質 A の短時間フーリエスペクトル 41a 上でフォルマント形状 50 の周波数

50a, 50bが500Hz及び3000Hzであり、声質Zの短時間フーリエスペクトル41a上でフォルマント形状50の周波数50a, 50bが400Hz及び4000Hzであり、かつ各合成音のナイキスト周波数が11025Hzである場合を想定して説明する。スペクトルモーフィング部205aは、まず、声質Aの短時間フーリエスペクトル41aの帯域 $f=0\sim500\text{Hz}$ が $0\sim(500+(400-500)*0.4)\text{Hz}$ となるように、帯域 $f=500\sim3000\text{Hz}$ が $(500+(400-500)*0.4)\sim(3000+(4000-3000)*0.4)\text{Hz}$ となるように、帯域 $f=3000\sim11025\text{Hz}$ が $(3000+(4000-3000)*0.4)\sim11025\text{Hz}$ となるように、声質Aの短時間フーリエスペクトル41aに対して周波数軸上の伸縮・移動を行う。これと同様に、スペクトルモーフィング部205aは、声質Zの短時間フーリエスペクトル41aの帯域 $f=0\sim400\text{Hz}$ が $0\sim(400+(500-400)*0.6)\text{Hz}$ となるように、帯域 $f=400\sim4000\text{Hz}$ が $(400+(500-400)*0.6)\sim(4000+(3000-4000)*0.6)\text{Hz}$ となるように、帯域 $f=4000\sim11025\text{Hz}$ が $(4000+(3000-4000)*0.6)\sim11025\text{Hz}$ となるように、声質Zの短時間フーリエスペクトル41aに対して周波数軸上の伸縮・移動を行う。その伸縮・移動の結果により生成された2つの短時間フーリエスペクトル41bにおいて、フォルマント形状50の周波数は周波数 f_1 , f_2 に揃えられた状態となる。

【0051】

次に、スペクトルモーフィング部205aは、このような周波数軸上の変形が行われた両短時間フーリエスペクトル41bのパワーを変形する。即ち、スペクトルモーフィング部205aは、声質Aの短時間フーリエスペクトル41bのパワーを60%に変換し、声質Zの短時間フーリエスペクトル41bのパワーを40%に変換する。そして、スペクトルモーフィング部205aは、上述のように、パワーが変換されたこれらの短時間フーリエスペクトルを重ね合わせる。

【0052】

図12は、パワーが変換された2つの短時間フーリエスペクトルを重ね合わせる様子を説明するための説明図である。

この図12に示すように、スペクトルモーフィング部205aは、パワーが変換された声質Aの短時間フーリエスペクトル41cと、同じくパワーが変換された声質Bの短時間フーリエスペクトル41cとを重ね合わせ、新たな短時間フーリエスペクトル41dを生成する。このとき、スペクトルモーフィング部205aは、互いの短時間フーリエスペクトル41cの上記周波数 f_1 , f_2 を一致させた状態で、両短時間フーリエスペクトル41cを重ね合わせる。

【0053】

そして、スペクトルモーフィング部205aは、上述のような短時間フーリエスペクトル41dの生成を、両合成音スペクトル41の時間軸アライメントされた時刻ごとに行う。その結果、声質Aの合成音スペクトル41と声質Zの合成音スペクトル41との音声モーフィング処理が4:6の割合で行われ、中間的合成音スペクトル42が生成されるのである。

【0054】

音声モーフィング部205の波形生成部205bは、上述のようにスペクトルモーフィング部205aにより生成された中間的合成音スペクトル42を、中間的合成音波形データ12に変換して、これをスピーカ107に出力する。その結果、スピーカ107から、中間的合成音スペクトル42に対応する合成音声が出力される。

このように、本実施の形態においても、実施の形態1と同様、声質の自由度が広く良い音質の合成音声をテキスト10から生成することができる。

【0055】

(変形例)

ここで、本実施の形態におけるスペクトルモーフィング部の動作に関する変形例につい

て説明する。

本変形例に係るスペクトルモーフィング部は、上述のように合成音スペクトル 4 1 からその形状の特徴を示すフォルマント形状 5 0 を抽出して用いることなく、音声合成 DB に予め格納されたスプライン曲線の制御点の位置を読み出して、そのスプライン曲線をフォルマント形状 5 0 の代わりに用いる。

【0 0 5 6】

即ち、各音声素片に対応するフォルマント形状 5 0 を、周波数対時間の 2 次元平面上の複数のスプライン曲線と見なし、そのスプライン曲線の制御点の位置を予め音声合成 DB に格納しておく。

このように、本変形例に係るスペクトルモーフィング部は、合成音スペクトル 4 1 からわざわざフォルマント形状 5 0 を抽出することをせず、音声合成 DB に予め格納されている制御点の位置が示すスプライン曲線を用いて時間軸及び周波数軸上の変換処理を行うため、上記変換処理を迅速に行うことができる。

【0 0 5 7】

なお、上述のようなスプライン曲線の制御点の位置ではなくフォルマント形状 5 0 そのものを、予め音声合成 DB 2 0 1 a ~ 2 0 1 z に格納しておいても良い。

【0 0 5 8】

(実施の形態 3)

図 1 3 は、本発明の実施の形態 3 に係る音声合成装置の構成を示す構成図である。

本実施の形態の音声合成装置は、実施の形態 1 の音声合成パラメタ値列 1 1 や、実施の形態 2 の合成音スペクトル 4 1 の代わりに音声波形を用い、この音声波形による音声モーフィング処理を行う。

【0 0 5 9】

このような音声合成装置は、複数の音声素片に関する音声素片データを蓄積する複数の音声合成 DB 3 0 1 a ~ 3 0 1 z と、1 つの音声合成 DB に蓄積された音声素片データを用いることにより、テキスト 1 0 に示される文字列に対応する合成音波形データ 6 1 を生成する複数の音声合成部 3 0 3 と、ユーザによる操作に基づいて声質を指定する声質指定部 1 0 4 と、複数の音声合成部 3 0 3 により生成された合成音波形データ 6 1 を用いて音声モーフィング処理を行い、中間的合成音波形データ 1 2 を出力する音声モーフィング部 3 0 5 と、中間的合成音波形データ 1 2 に基づいて合成音声出力するスピーカ 1 0 7 とを備えている。

【0 0 6 0】

複数の音声合成 DB 3 0 1 a ~ 3 0 1 z のそれぞれが蓄積する音声素片データの示す声質は、実施の形態 1 の音声合成 DB 1 0 1 a ~ 1 0 1 z と同様、異なっている。また、本実施の形態における音声素片データは、音声波形の形式で表現されている。

複数の音声合成部 3 0 3 は、それぞれ上述の音声合成 DB と一対一に対応付けられている。そして、各音声合成部 3 0 3 は、テキスト 1 0 を取得して、テキスト 1 0 に示される文字列を音素情報に変換する。さらに、音声合成部 3 0 3 は、対応付けられた音声合成 DB の音声素片データから適切な音声素片に関する部分を抜き出して、抜き出した部分の結合と変形を行うことにより、先に生成した音素情報に対応する音声波形たる合成音波形データ 6 1 を生成する。

【0 0 6 1】

声質指定部 1 0 4 は、実施の形態 1 と同様、ユーザによる操作に基づき、何れの合成音波形データ 6 1 を用い、その合成音波形データ 6 1 に対してどのような割合で音声モーフィング処理を行うかを音声モーフィング部 3 0 5 に指示する。さらに、声質指定部 1 0 4 はその割合を時系列に沿って変化させる。

本実施の形態における音声モーフィング部 3 0 5 は、複数の音声合成部 3 0 3 から出力される合成音波形データ 6 1 を取得して、その中間的性質を持つ中間的合成音波形データ 1 2 を生成して出力する。

【0 0 6 2】

図14は、本実施の形態における音声モーフィング部305の処理動作を説明するための説明図である。

本実施の形態における音声モーフィング部305は波形編集部305aを備えている。

この波形編集部305aは、声質指定部104により指定された少なくとも2つの合成音波形データ61と割合とを特定し、それらの合成音波形データ61から、その割合に応じた中間的合成音波形データ12を生成する。

【0063】

即ち、波形編集部305aは、複数の合成音波形データ61から、声質指定部104により指定された2つ以上の合成音波形データ61を選択する。そして、波形編集部305aは、声質指定部104により指定された割合に応じ、その選択した合成音波形データ61のそれぞれに対して、例えば各音声の各サンプリング時点におけるピッチ周波数や振幅、各音声における各有声区間の継続時間長などを変形する。波形編集部305aは、そのように変形された合成音波形データ61を重ね合わせることで、中間的合成音波形データ12を生成する。

【0064】

スピーカ107は、このように生成された中間的合成音波形データ12を波形編集部305aから取得して、その中間的合成音波形データ12に対応する合成音声を出力する。

このように、本実施の形態においても、実施の形態1又は2と同様、声質の自由度が広く良い音質の合成音声をテキスト10から生成することができる。

【0065】

(実施の形態4)

図15は、本発明の実施の形態4に係る音声合成装置の構成を示す構成図である。

本実施の形態の音声合成装置は、出力する合成音声の声質に応じた顔画像を表示するのであって、実施の形態1に含まれる構成要素と、複数の顔画像に関する画像情報を蓄積する複数の画像DB401a～401zと、これらの画像DB401a～401zに蓄積される顔画像の情報をを用いて画像モーフィング処理を行い、中間的顔画像データ12pを出力する画像モーフィング部405と、画像モーフィング部405から中間的顔画像データ12pを取得して、その中間的顔画像データ12pに応じた顔画像を表示する表示部407とを備えている。

【0066】

画像DB401a～401zのそれぞれが蓄積する画像情報の示す顔画像の表情は異なっている。例えば、怒っている声質の音声合成DB101aに対応する画像DB401aには、怒っている表情の顔画像に関する画像情報が蓄積されている。また、画像DB401a～401zに蓄積されている顔画像の画像情報には、顔画像の眉及び口の端や中央、目の中心点など、この顔画像の表す表情の印象をコントロールするための特徴点が付加されている。

【0067】

画像モーフィング部405は、声質指定部104により指定された各合成音声パラメタ値列102のそれぞれの声質に対応付けされた画像DBから画像情報を取得する。そして、画像モーフィング部405は、取得した画像情報を用いて、声質指定部104により指定された割合に応じた画像モーフィング処理を行う。

具体的に、画像モーフィング部405は、取得した一方の画像情報により示される顔画像の特徴点の位置が、声質指定部104により指定された割合だけ、取得した他方の画像情報により示される顔画像の特徴点の位置に変位するように、その一方の顔画像をワーピングし、これと同様に、その他方の顔画像の特徴点の位置を、声質指定部104により指定された割合だけ、その一方の顔画像の特徴点の位置に変位するように、その他方の顔画像をワーピングする。そして、画像モーフィング部405は、ワーピングされたそれぞれの顔画像を、声質指定部104により指定された割合に応じてクロスディゾルブすることで、中間的画像データ12pを生成する。

【0068】

これにより本実施の形態では、例えばエージェントの顔画像と合成音声の声質の印象を常に一致させることができる。即ち、本実施の形態の音声合成装置は、エージェントの平常声と怒り声の間の音声モーフィングを行って、少しだけ怒った声質の合成音声を生成するときには、音声モーフィングと同様の比率でエージェントの平常顔画像と怒り顔画像の間の画像モーフィングを行い、エージェントのその合成音声に適した少しだけ怒った顔画像を表示する。言い換えれば、感情を持つエージェントに対してユーザが感じる聴覚的印象と、視覚的印象を一致させることができ、エージェントの提示する情報の自然性を高めることができる。

【0069】

図16は、本実施の形態の音声合成装置の動作を説明するための説明図である。

例えば、ユーザが声質指定部104を操作することにより、図3に示すディスプレイ上の指定アイコン104iを、声質アイコン104Aと声質アイコン104Zを結ぶ線分を4:6に分割する位置に配置すると、音声合成装置は、スピーカ107から出力される合成音声は10%だけ声質A寄りになるように、その4:6の割合に応じた音声モーフィング処理を声質A及び声質Zの音声合成パラメタ値列11を用いて行い、声質A及び声質Bの中間的な声質xの合成音声を出力する。これと同時に、音声合成装置は、上記割合と同じ4:6の割合に応じた画像モーフィング処理を、声質Aに対応付けられた顔画像P1と、声質Zに対応付けられた顔画像P2とを用いて行い、これらの画像の中間的な顔画像P3を生成して表示する。ここで、音声合成装置は、画像モーフィングするときには、上述のように、顔画像P1の眉や口の端などの特徴点の位置を、顔画像P2の眉や口の端などの特徴点の位置に向けて40%の割合で変化するように、その顔画像P1をワーピングし、これと同様に、顔画像P2の特徴点の位置を、顔画像P1の特徴点の位置に向けて60%の割合で変化するように、その顔画像P2をワーピングする。そして、画像モーフィング部405は、ワーピングされた顔画像P1に対して60%の割合で、ワーピングされた顔画像P2に対して40%の割合でクロスディゾルブし、その結果、顔画像P3を生成する。

【0070】

このように、本実施の形態の音声合成装置は、スピーカ107から出力する合成音声の声質が「怒っている」とときには、「怒っている」様子の顔画像を表示部407に表示し、声質が「泣いている」とときには、「泣いている」様子の顔画像を表示部407に表示する。さらに、本実施形態の音声合成装置は、その声質が「怒っている」と「泣いている」との中間的なものであるときには、「怒っている」顔画像と「泣いている」顔画像の中間的な顔画像を表示するとともに、その声質が「怒っている」ものから「泣いている」ものへと時間的に変化するときには、中間的な顔画像をその声質に一致させて時間的に変化させる。

【0071】

なお、画像モーフィングは他にも様々な方法によって可能であるが、元となる画像の間の比率を指定することで目的の画像が指定できる方法であれば、どんなものを用いてもよい。

【産業上の利用可能性】

【0072】

本発明は、声質の自由度が広く良い音質の合成音声テキストデータから生成することができるという効果を有し、ユーザに対して感情を表す合成音声出力する音声合成装置などに適用することができる。

【図面の簡単な説明】

【0073】

【図1】 本発明の実施の形態1に係る音声合成装置の構成を示す構成図である。

【図2】 同上の音声合成部の動作を説明するための説明図である。

【図3】 同上の声質指定部のディスプレイが表示する画面の一例を示す画面表示図である。

【図 4】 同上の声質指定部のディスプレイが表示する他の画面の一例を示す画面表示図である。

【図 5】 同上の音声モーフィング部の処理動作を説明するための説明図である。

【図 6】 同上の音声素片と HMM 音素モデルの一例を示す例示図である。

【図 7】 同上の変形例に係る音声合成装置の構成を示す構成図である。

【図 8】 本発明の実施の形態 2 に係る音声合成装置の構成を示す構成図である。

【図 9】 同上の音声モーフィング部の処理動作を説明するための説明図である。

【図 10】 同上の声質 A 及び声質 Z の合成音スペクトルと、それらに対応する短時間フーリエスペクトルとを示す図である。

【図 11】 同上のスペクトルモーフィング部が両短時間フーリエスペクトルを周波数軸上で伸縮する様子を説明するための説明図である。

【図 12】 同上のパワーが変換された 2 つの短時間フーリエスペクトルを重ね合わせる様子を説明するための説明図である。

【図 13】 本発明の実施の形態 3 に係る音声合成装置の構成を示す構成図である。

【図 14】 同上の音声モーフィング部の処理動作を説明するための説明図である。

【図 15】 本発明の実施の形態 4 に係る音声合成装置の構成を示す構成図である。

【図 16】 同上の音声合成装置の動作を説明するための説明図である。

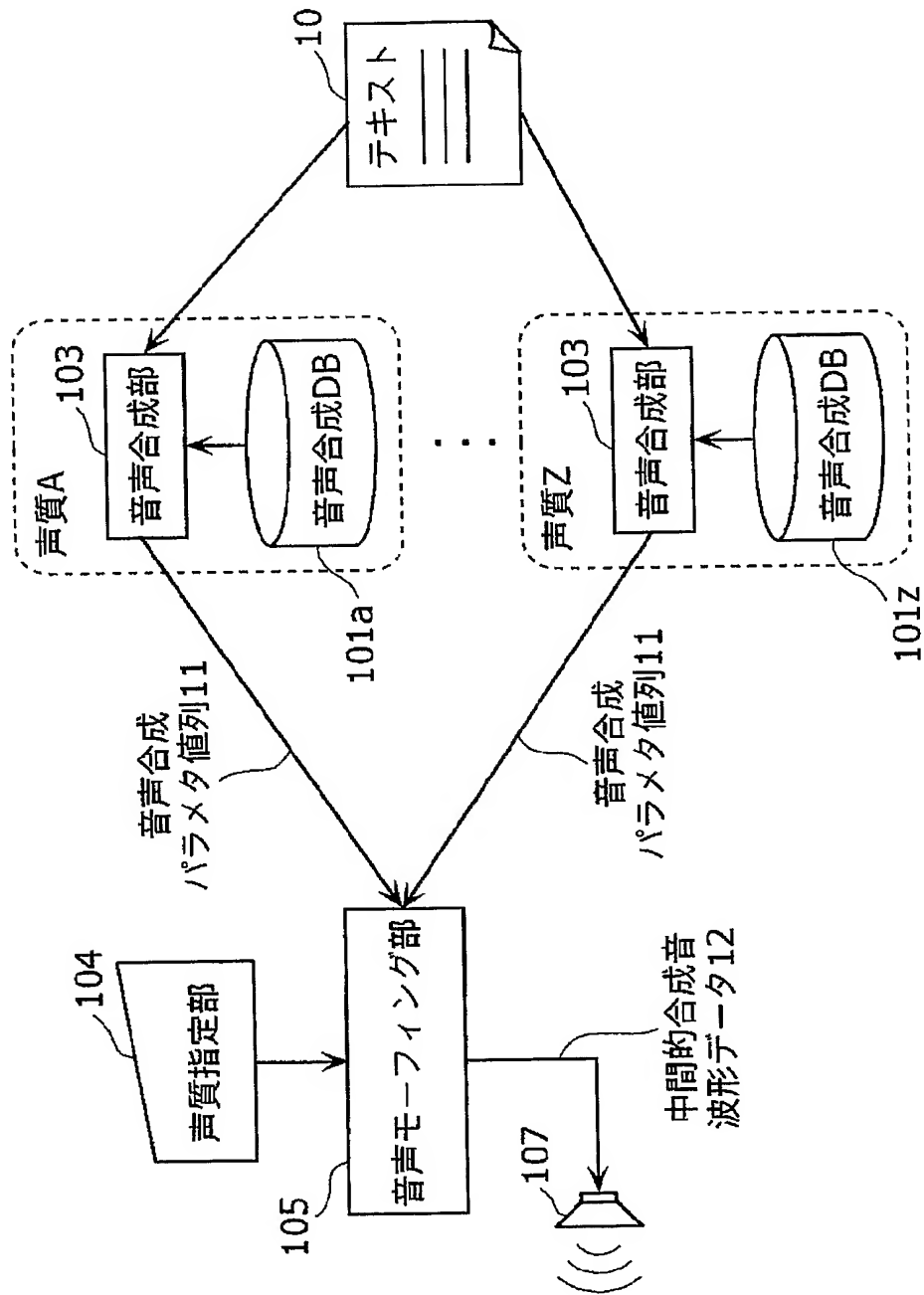
【符号の説明】

【0074】

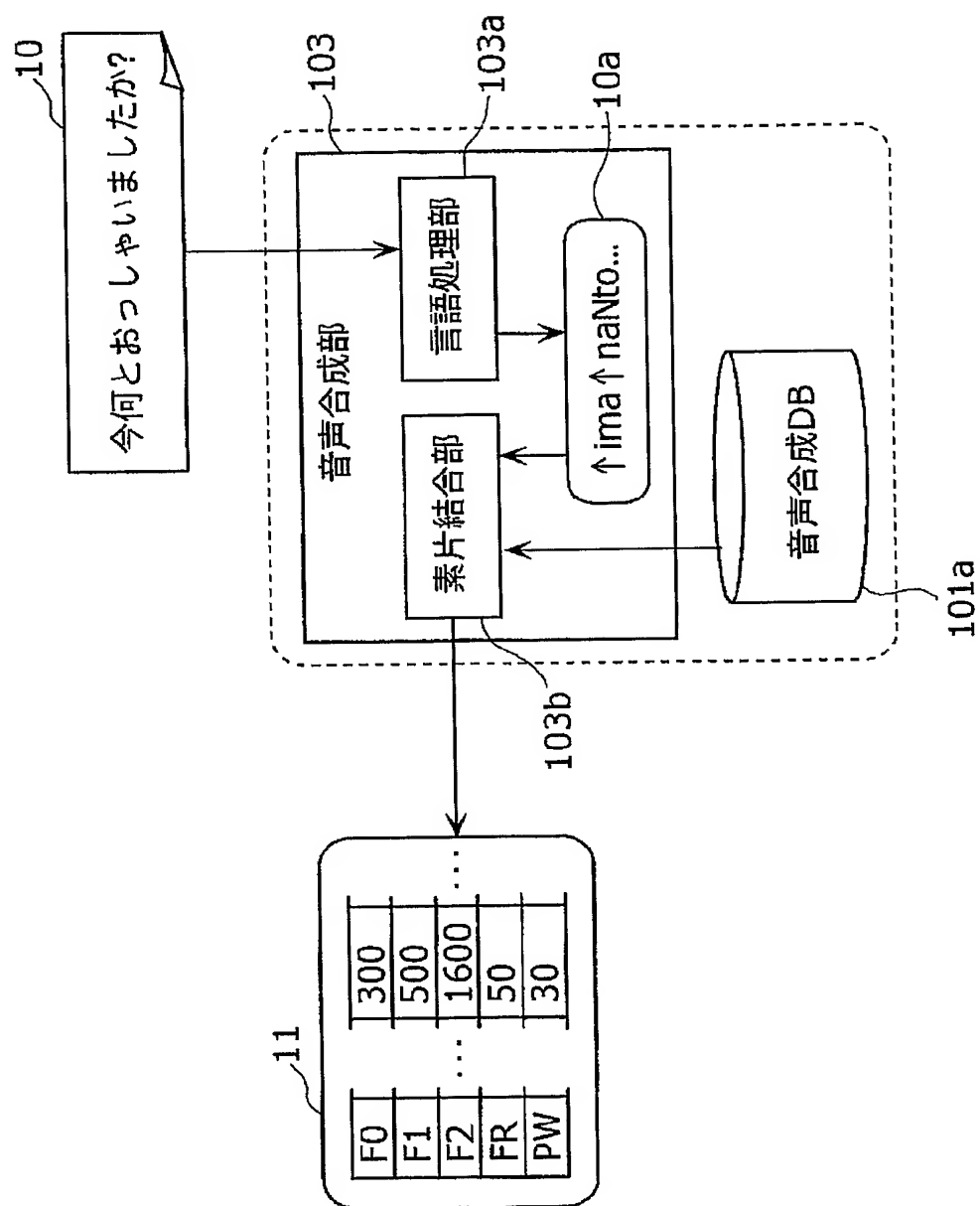
- 10 テキスト
- 10a 音素情報
- 11 音声合成パラメタ値列
- 12 中間的合成音波形データ
- 12p 中間的画像データ
- 13 中間的音声合成パラメタ値列
- 30 音声素片
- 31 音素モデル
- 32 最尤パスの形状
- 41 合成音スペクトル
- 42 中間的合成音スペクトル
- 50 フォルマント形状
- 50a, 50b 周波数
- 51 フーリエスペクトル分析窓
- 61 合成音波形データ
- 101a ~ 101z 音声合成 DB
- 103 音声合成部
- 103a 言語処理部
- 103b 素片結合部
- 104 声質指定部
- 104A, 104B, 104Z 声質アイコン
- 104i 指定アイコン
- 105 音声モーフィング部
- 105a パラメタ中間値計算部
- 105b 波形生成部
- 106 中間的合成音波形データ
- 107 スピーカ
- 203 音声合成部
- 201a ~ 201z 音声合成 DB
- 205 音声モーフィング部
- 205a スペクトルモーフィング部

2 0 5 b 波形生成部
3 0 3 音声合成部
3 0 1 a ~ 3 0 1 z 音声合成 D B
3 0 5 音声モーフィング部
3 0 5 a 波形編集部
4 0 1 a ~ 4 0 1 z 画像 D B
4 0 5 画像モーフィング部
4 0 7 表示部
P 1 ~ P 3 顔画像

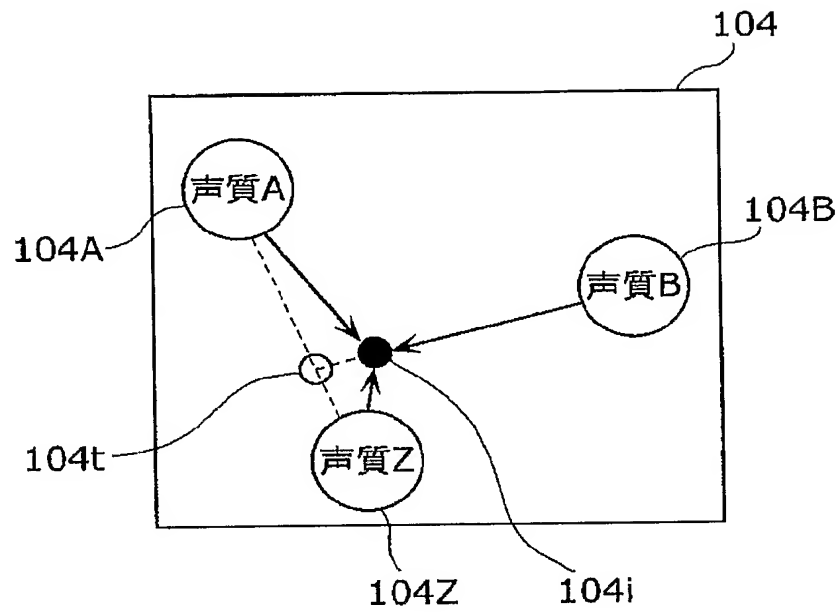
【書類名】 図面
【図 1】



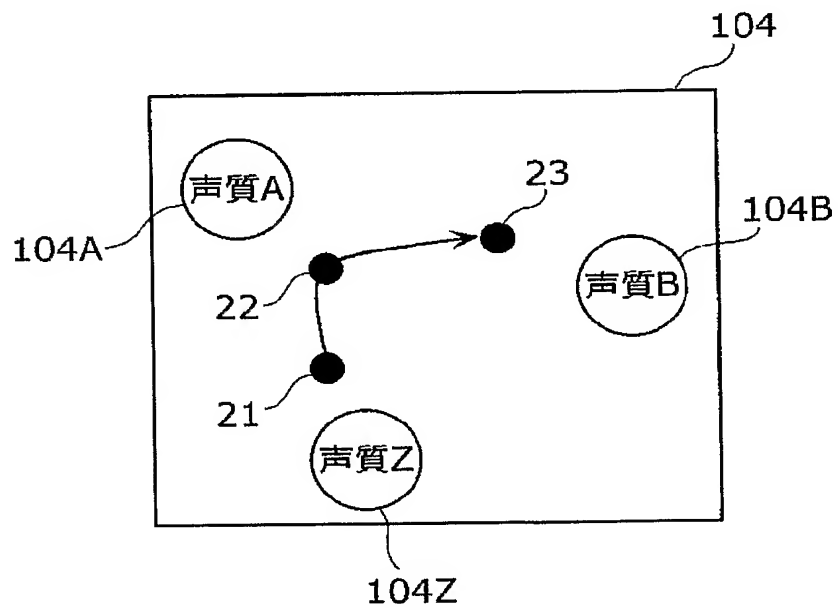
【図 2】



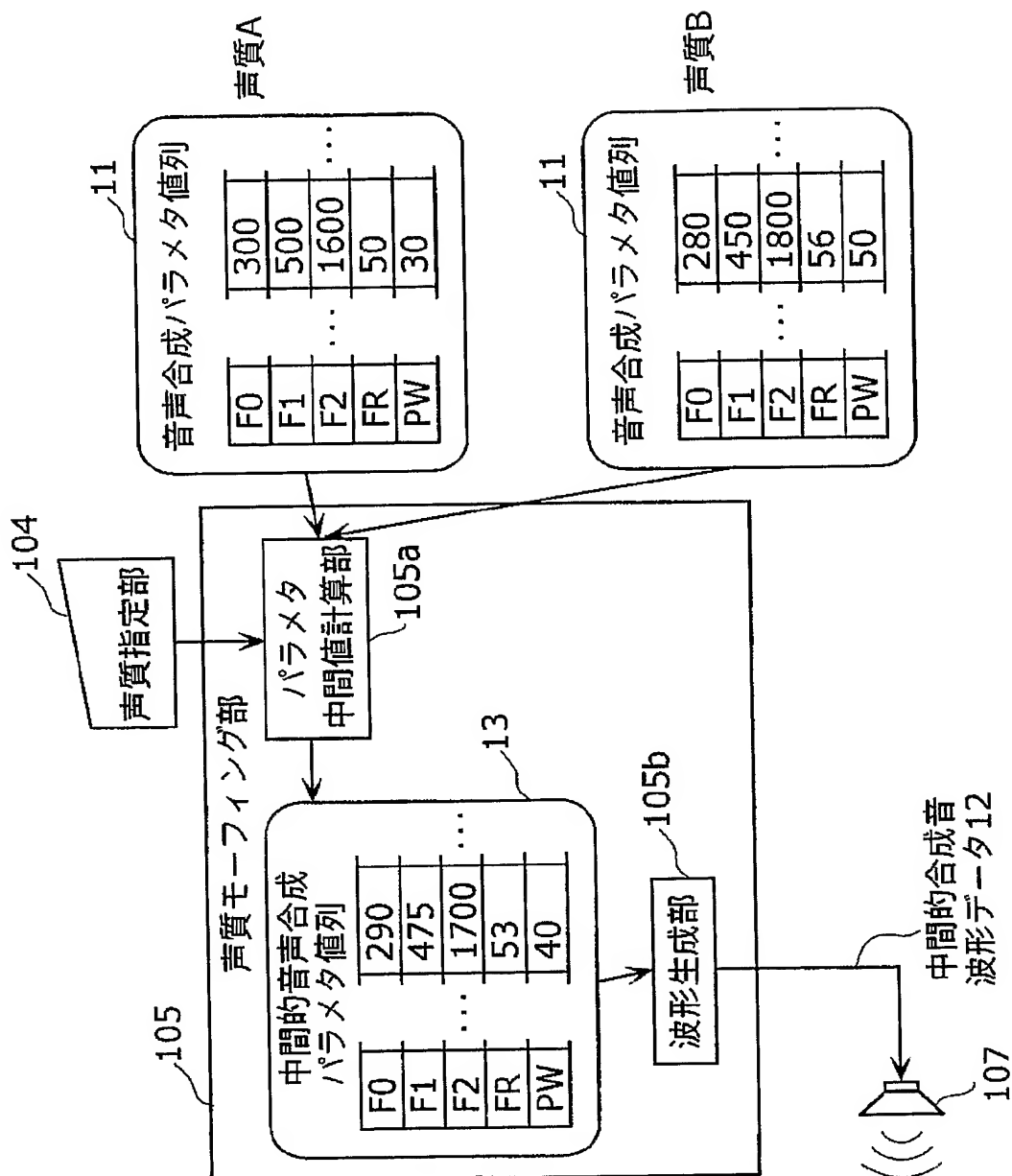
【図 3】



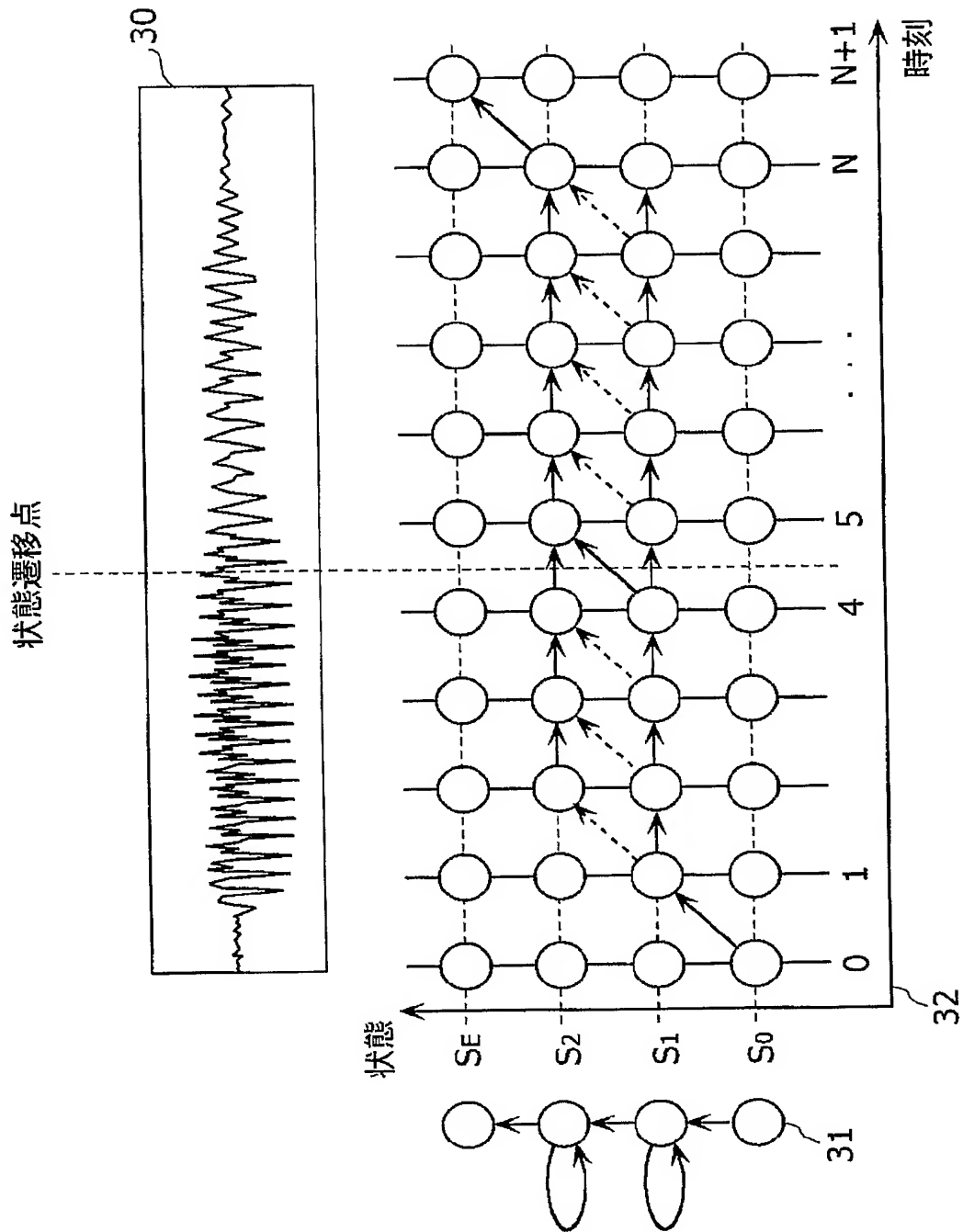
【図 4】



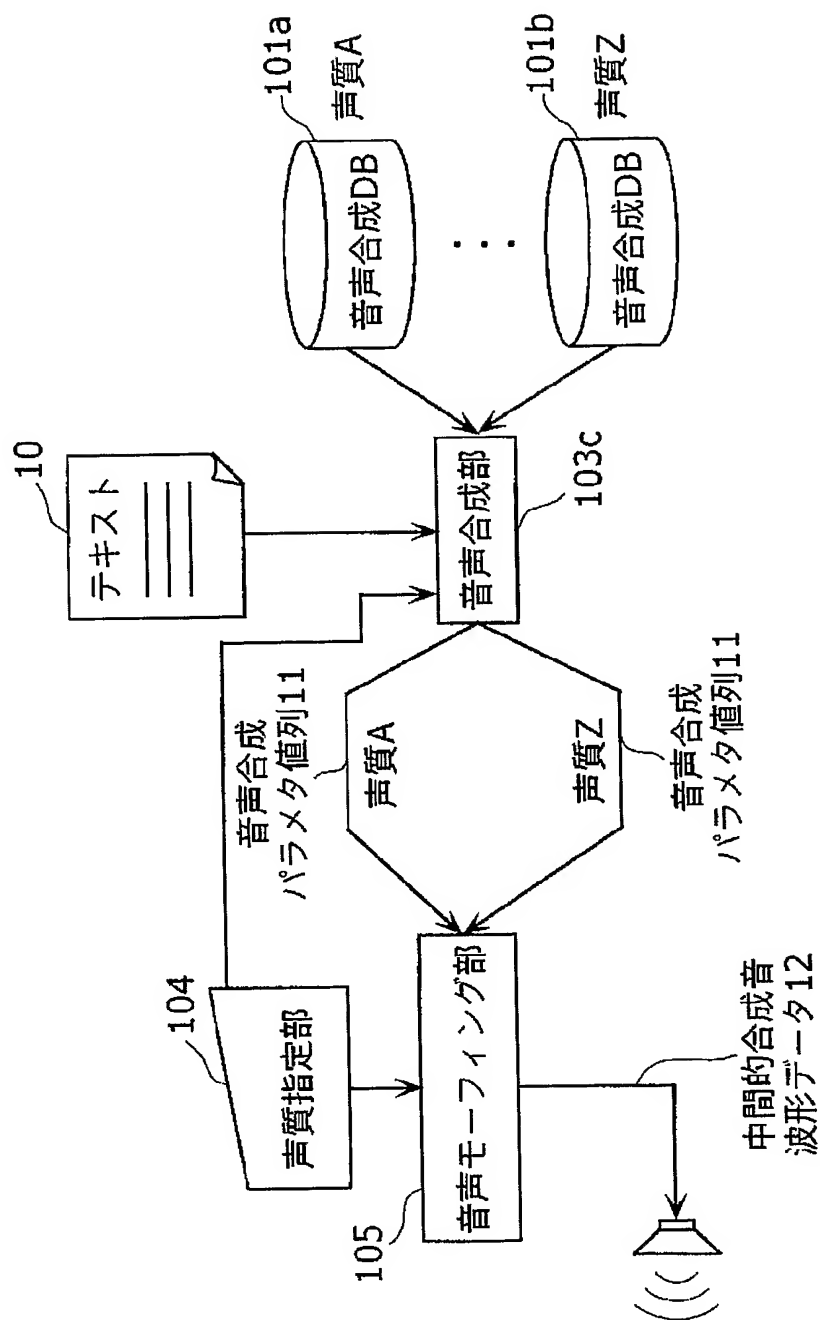
【図 5】



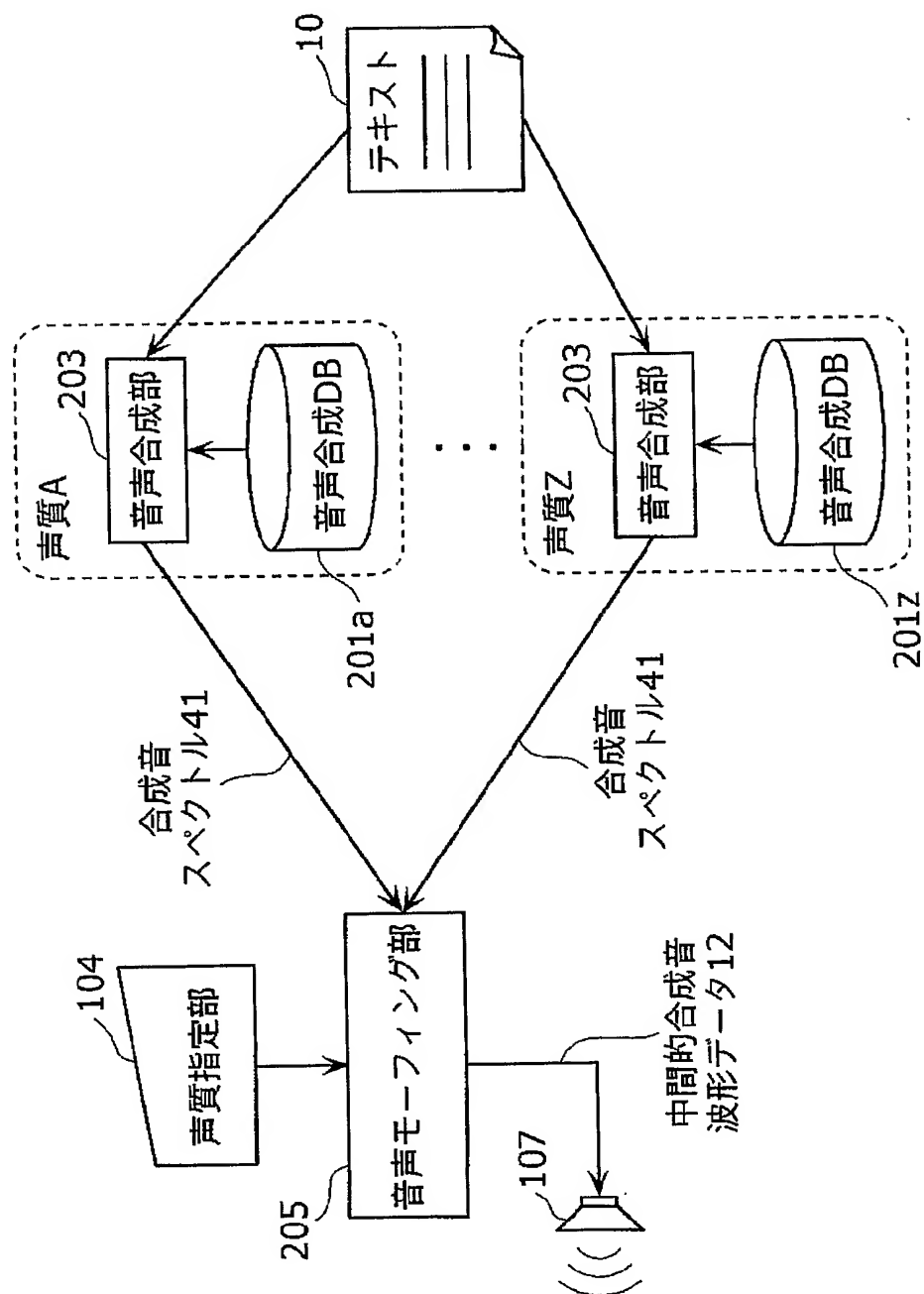
【図 6】



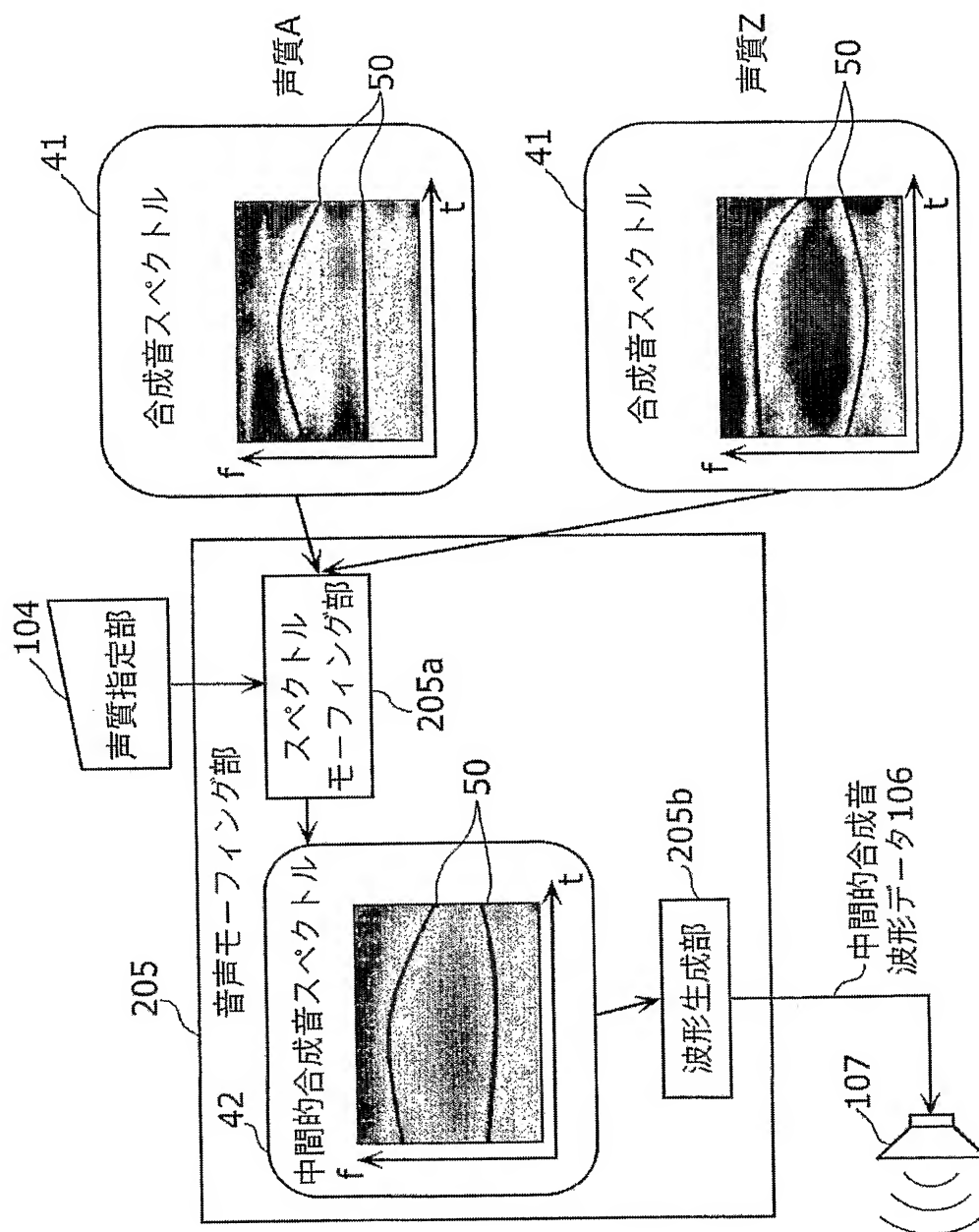
【図 7】



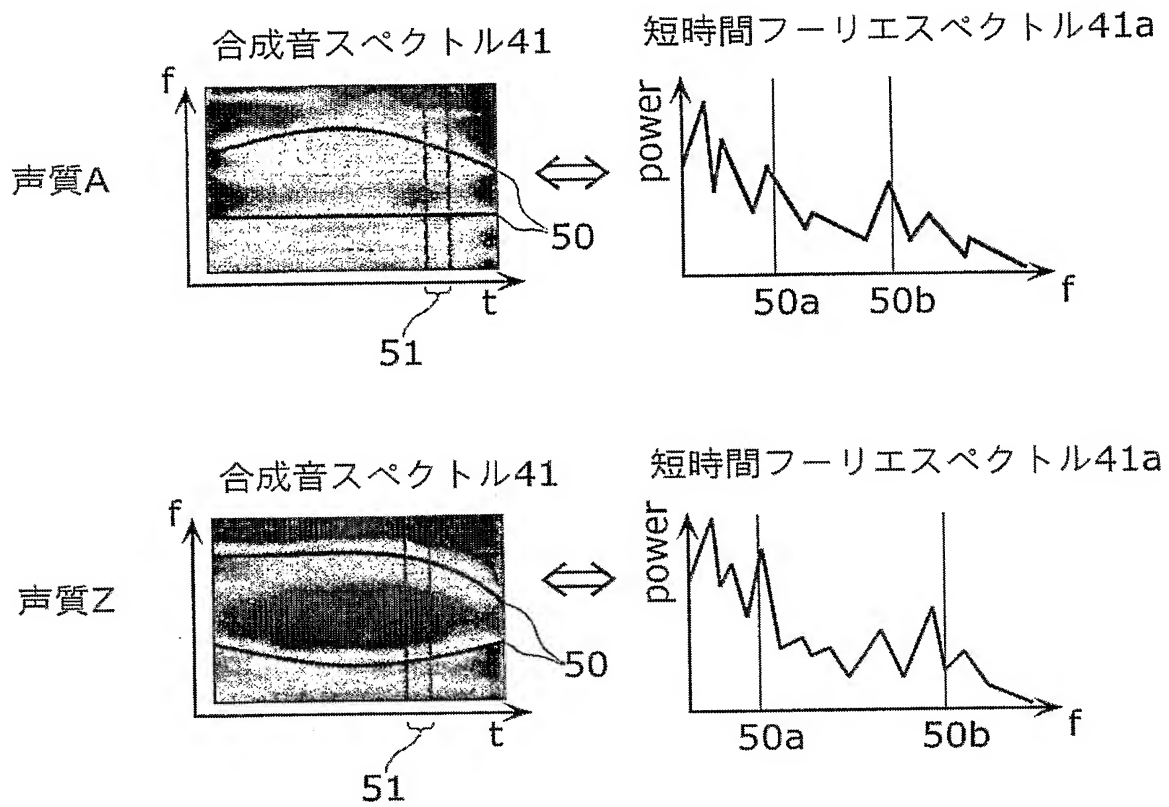
【図 8】



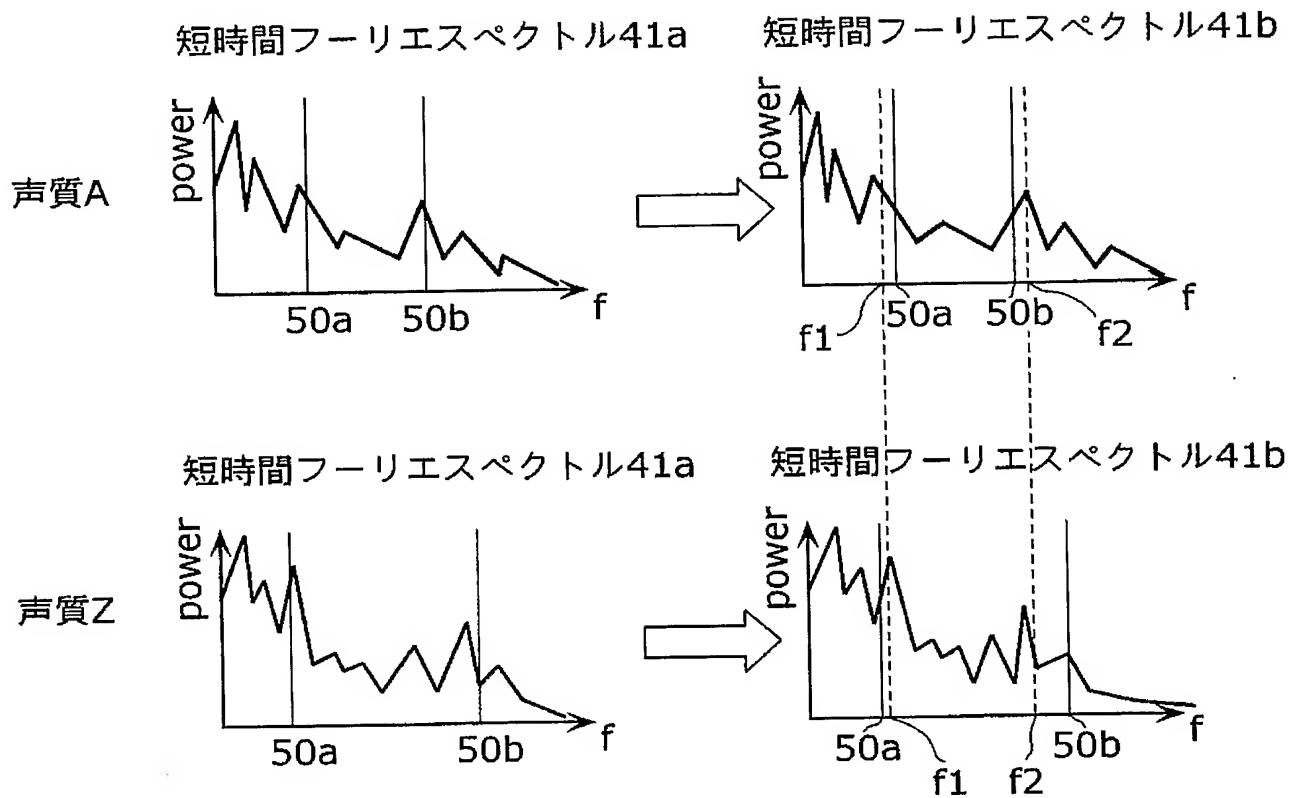
【図 9】



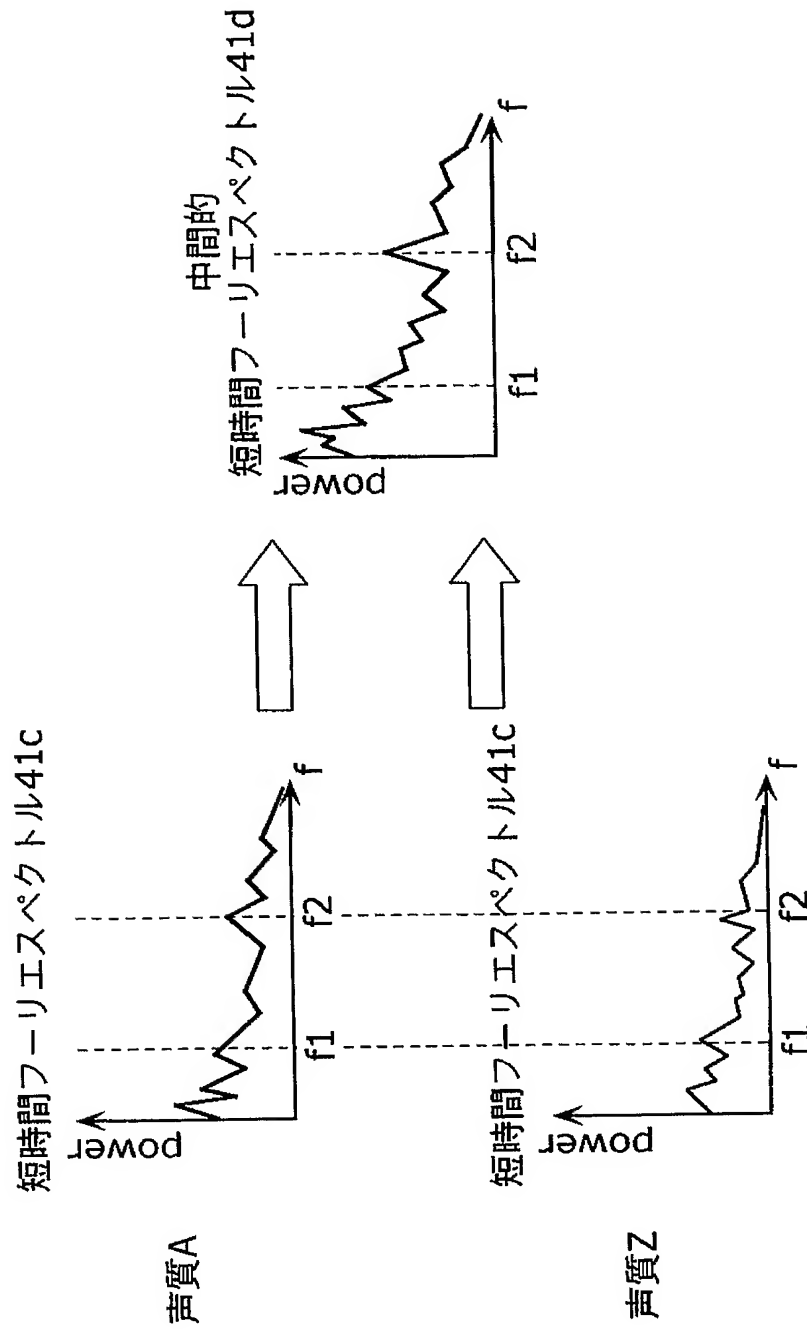
【図 10】



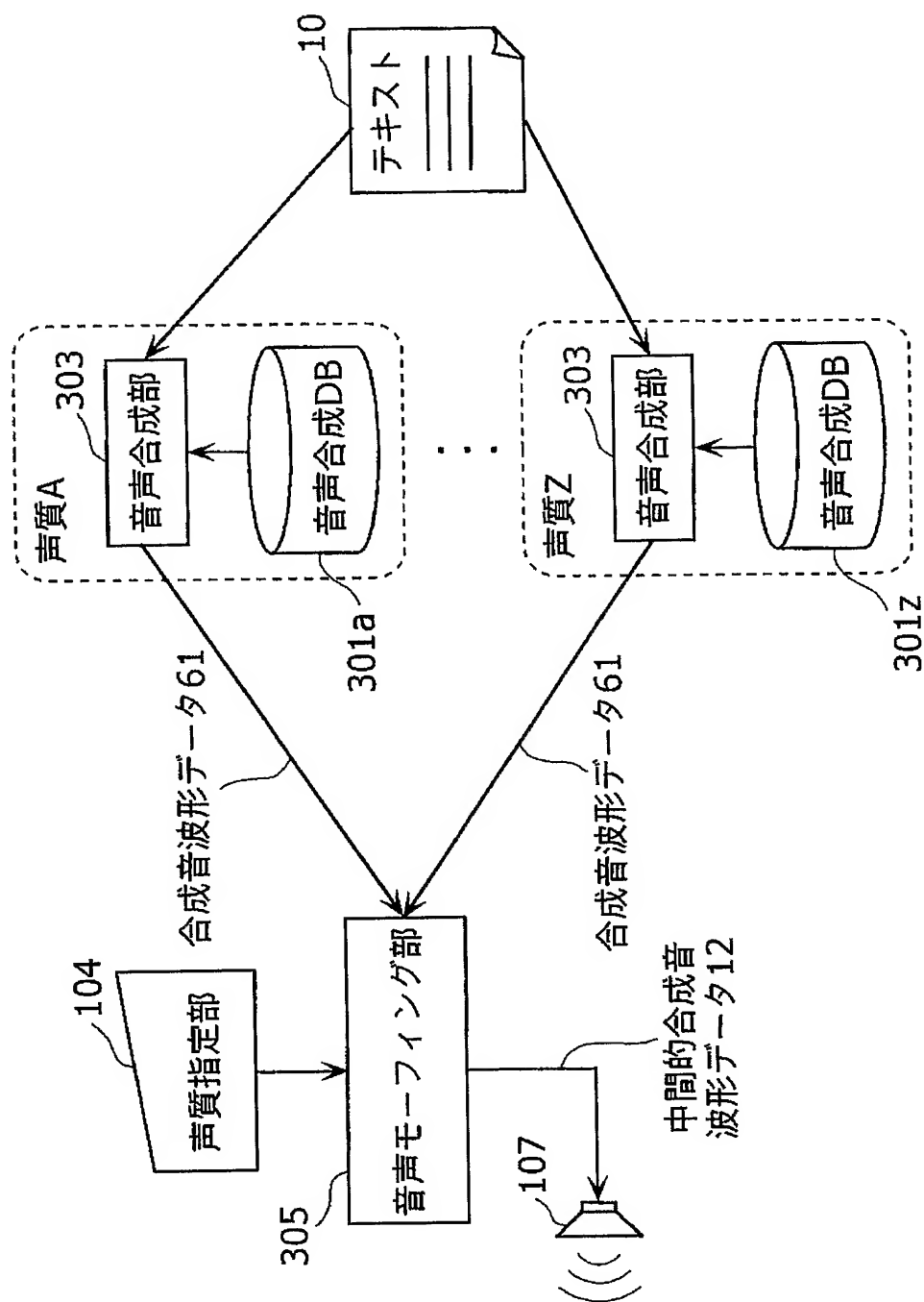
【図 11】



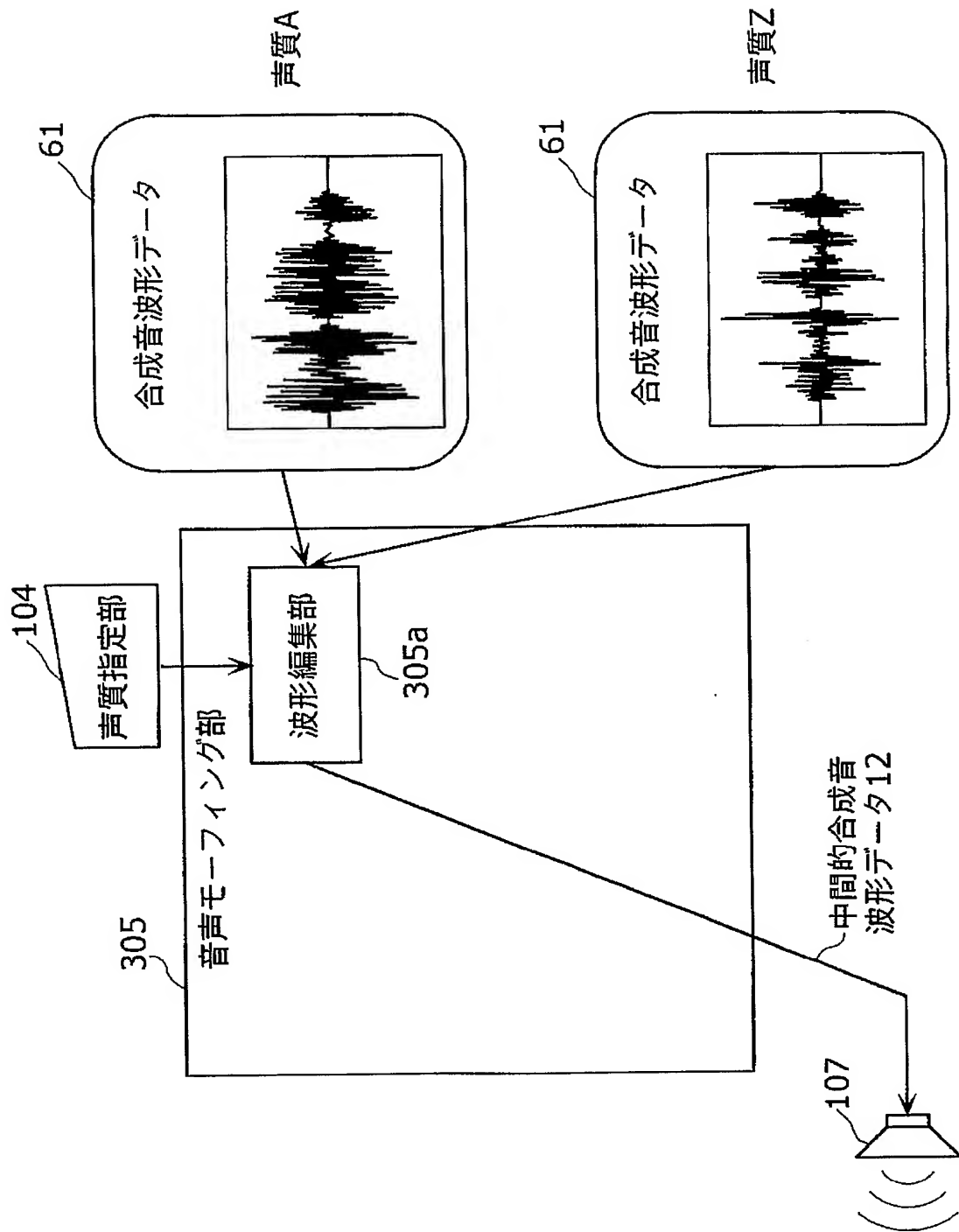
【図 12】



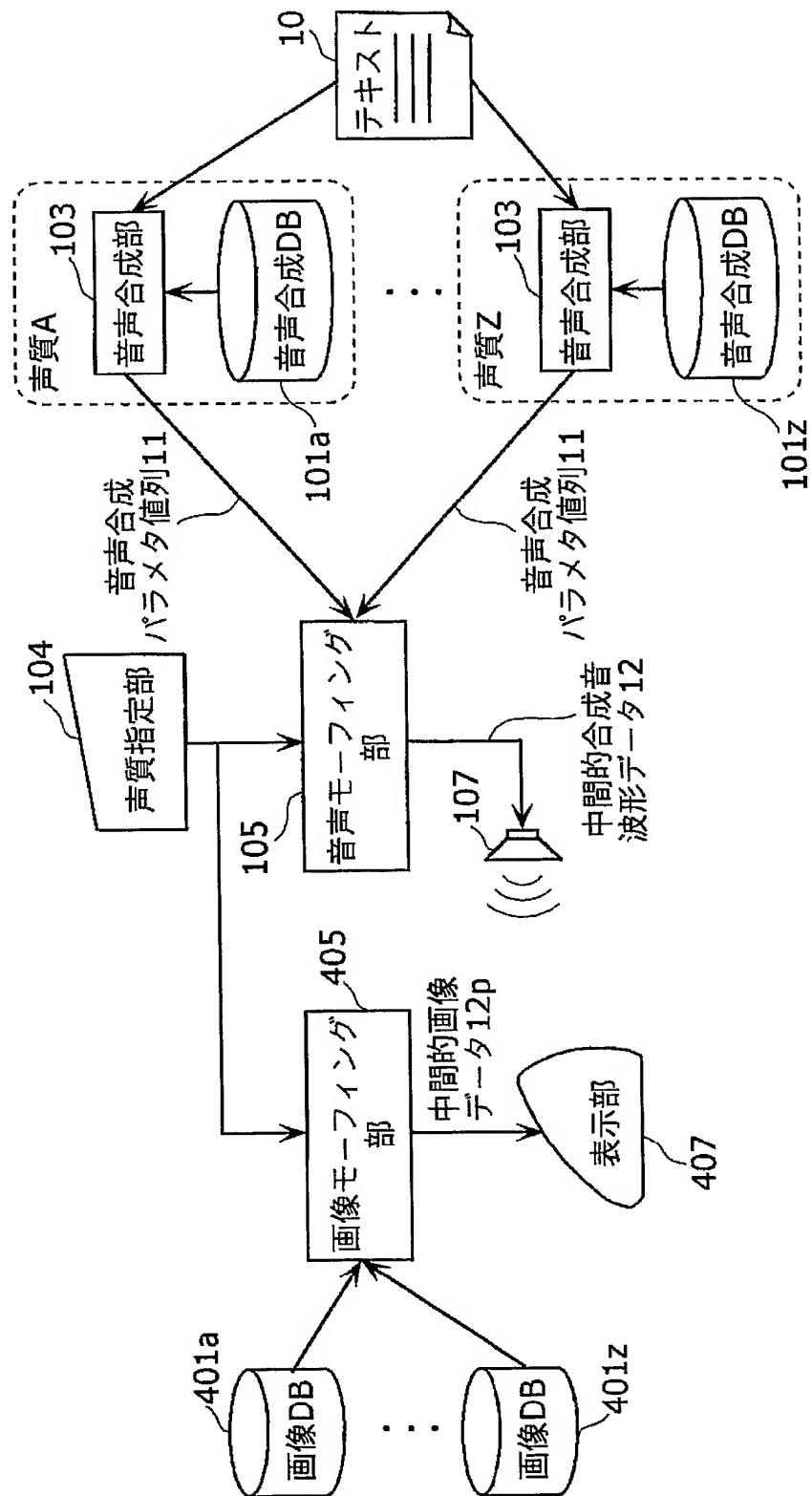
【図 13】



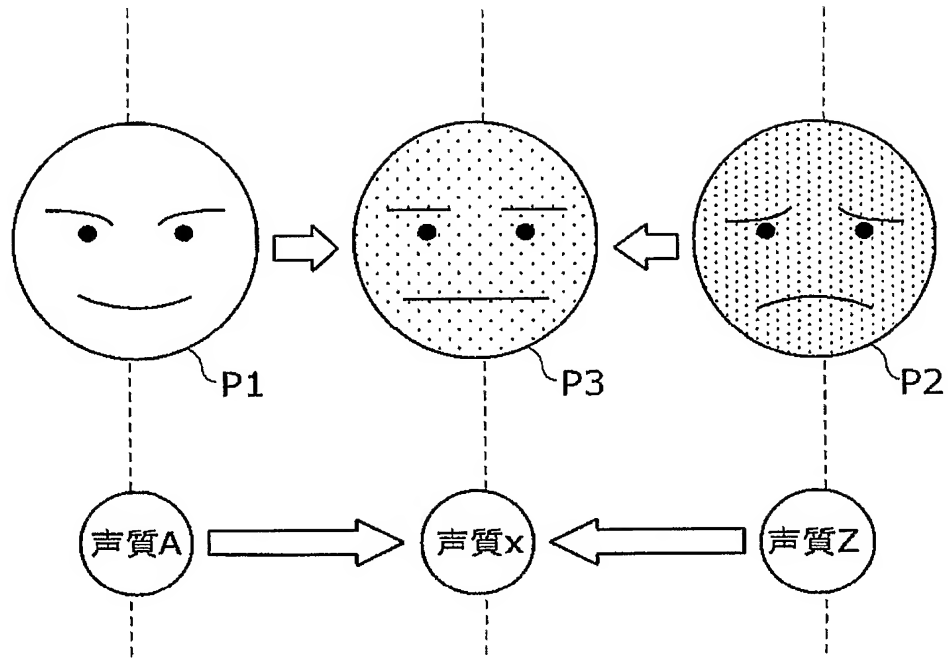
【図 14】



【図 15】



【図 16】



【書類名】 要約書

【要約】

【課題】 声質の自由度が広く良い音質の合成音声テキストデータから生成する音声合成装置を提供することを目的とする。

【解決手段】 音声合成装置は、音声合成DB101a、101zと、テキスト10を取得するとともに、音声合成DB101aから、テキスト10に含まれる文字に対応した声質Aの音声合成パラメタ値列11を生成する音声合成部103と、音声合成DB101zから、テキスト10に含まれる文字に対応した声質Zの音声合成パラメタ値列11を生成する音声合成部103と、声質A及び声質Zの音声合成パラメタ値列11から、テキスト10に含まれる文字に対応した、声質A及び声質Zの中間的な声質の合成音声を示す中間的音声合成パラメタ値列13を生成する音声モーフィング部105と、生成された中間的音声合成パラメタ値列13をその合成音声に変換して出力するスピーカ107とを備える。

【選択図】 図1

認定・付加情報

特許出願の番号	特願 2 0 0 4 - 0 1 8 7 1 5
受付番号	5 0 4 0 0 1 3 3 2 7 7
書類名	特許願
担当官	第八担当上席 0 0 9 7
作成日	平成 1 6 年 1 月 2 8 日

< 認定情報・付加情報 >

【提出日】 平成 16 年 1 月 27 日

特願 2 0 0 4 - 0 1 8 7 1 5

ページ : 1/E

出 願 人 履 歴 情 報

識別番号

[0 0 0 0 0 5 8 2 1]

1. 変更年月日

1 9 9 0 年 8 月 2 8 日

[変更理由]

新規登録

住 所

大阪府門真市大字門真 1 0 0 6 番地

氏 名

松下電器産業株式会社